A THEORY OF SPEECH PERCEPTION IN NORMAL AND HEARING-IMPAIRED EARS

BY

RIYA OMPRAKASH SINGH

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Electrical and Computer Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2010

Urbana, Illinois

Adviser:

Associate Professor Jont B. Allen

ABSTRACT

This research investigates the longstanding problem of understanding human speech perception. We aim to study speech perception and decode perceptual cues in speech by conducting psychoacoustic experiments on several subjects by presenting them with nonsense consonant-vowel (CV) syllables in various kinds of masking noise at different signal-to-noise ratios (SNRs). Our research with a large number of normal hearing (NH) listeners shows that NH speech perception is deterministic and the error is essentially zero when the main perceptual feature (or event) of the utterance is audible. With the perceptual CV cues precisely known, one can predict how an average normal hearing (ANH) listener would behave in a certain type and degree of masking noise. The next major goal of the current research is to characterize hearing-impaired (HI) ears by using our knowledge of specific consonant speech cues in ANH ears, thus quantifying how the HI ears differ from ANH ears in their use of acoustic cues. Our analysis shows that HI ears may have poor temporal and/or frequency resolution, because of which they are unable to hear only a few consonants, yet they can hear the rest. We argue that it is necessary to measure this consonant dependence in order to design a more sensitive hearing aid fitting technique, and no other clinical measure used currently (audiometry, average speech recognition scores, speech in noise tests) is useful in characterizing speech-loss, in HI ears. We measured 46 HI ears with our CV discrimination test using the current hearing aid amplification

technique NAL-R; the results show that though NAL-R improves the average score, it degrades a few consonants under certain circumstances. This research also addresses the important issue of cochlear dead regions, which are places along the basilar membrane of the cochlea where the inner hair cells are degenerate. We propose a new method to diagnose dead regions based on comodulation masking release. This project extends our effort to achieve a fundamental insight into the nature of both ANH and HI speech perception, enabling the design of hearing aids that are functionally useful in high ambient noise and that help make audible the sounds that the HI ear could not hear previously, without affecting the sounds that they can hear.

To my mother, who means the world to me.

I think I must tell you this more often - "Love you Mommy."

ACKNOWLEDGMENTS

I first wish to thank Professor Richard Sproat, who got me into UIUC. The thesis would not have been possible without my adviser, Dr. Jont Allen, who gave me a chance to work with his group when I was adviser-less after my first semester. I thank him for all the support and suggestions and for being my "mom" during my two-year stint at UIUC. Thanks to Kunal, for being my best friend for all these years and being a part of so many of my happy memories. Thanks to my family - my brother (who enjoys my successes more than I do), my father (who keeps wanting to proofread my papers - here Dad, the whole thesis for you to read!), my mother (for well ... everything).

A number of people have contributed substantially to this work. Special thanks to Woojae for collecting the data so efficiently and being my best buddy and supporter. Thanks to all members of the Human Speech Recognition Group, especially Anjali, Abhinauv, Roger and Andrea for their many helpful suggestions and encouragement. I sincerely thank everyone at the ECE Publications Office, who have been so kind and helpful to me.

There have been times when I have been lost and unsure about my work and my goals. Thanks to Abhradeep, Daksh, Vineet and Parikshit for talking with me during those times and helping me clear my mind. A big thank you to Vinni - for the energy boosters, the numerous trips together, the Tuesday lunch dates, the (usually lovely) Friday evenings, and of course the chicken biryani and the gulab jamuns! I also want to thank Pari Bhaiya for being so

kind and nice and fun and for the numerous treats. (And yup, I'll give you a treat soon upon thesis completion.)

And finally a big thanks to Champaign! I have complained a lot about you - your moody weather swings, your cold windy nights, your hot humid summers, your loneliness - but I have met some really special people and have enjoyed my time here. Thanks Chambana - you have been good to me.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

According to the National Institute on Deafness and Other Communication Disorders (NIDCD), approximately 30% of the United States population above the age of 65 and about 47% of the population above the age of 75 have hearing loss. This percentage is increasing as life expectancy increases. Moreover, results of newborn hearing screening estimate that approximately 0.2∼0.3% of children in the United States are born deaf or hard of hearing. Such congenital hearing disorders severely limit the child's ability to learn important communication skills since the first few years of life are the most critical for developing language [42]. The most common type of hearing loss in these populations is called *sensorineural hearing loss* (SNHL) which is mainly caused by damage to cochlea hair cells and/or the auditory nerve.

1.1 Problem statement and approach

The goal of this research is to provide a deep understanding of how HI ears perceive speech. For understanding HI speech perception, it is critical to first understand how normals decode speech. Having a theory of human speech recognition (HSR) is critical for the development of new hearing aids. To research this longstanding problem, we have measured a large number of normal listeners' responses to individual consonant-vowel (CV) syllables in noise [34, 33, 57]. We then correlate the confusions with the acoustic

cues of the utterances to derive the perceptual features or *events*. Next we characterize HI ears by using our knowledge of specific consonant speech cues in average normal hearing (ANH) ears, thus quantifying how the HI ears differ from ANH ears in their use of acoustic cues.

1.2 Psychoacoustics

Psychoacoustics is the study of subjective human perception of sounds. It is the transformation from the physical domain that contains the acoustic features (or physical variables ϕ) to the psychological domain that contains the perceptual cues or *events* (psychological variables ψ) via the listener, as shown in Fig. 1.1. An example is acoustic intensity (the ϕ or physical intensity) and loudness (the ψ or psychoacoustic intensity). In the case of speech perception, we treat physical variables as analog (continuous) and psychophysical variables as discrete, as in the case of events. Event (perceptual feature) is the ψ correlate to the ϕ acoustic feature.

Figure 1.1: Psychoacoustics transformation: The basic model of an observer with the physical variables ϕ on the left and the psychological variables ψ on the right.

In 1948, Claude Shannon [60] proposed the information-theoretic model of the modern communication channel, the basic elements of which are shown in Fig. 1.2.

We use a similar model for understanding human speech perception as

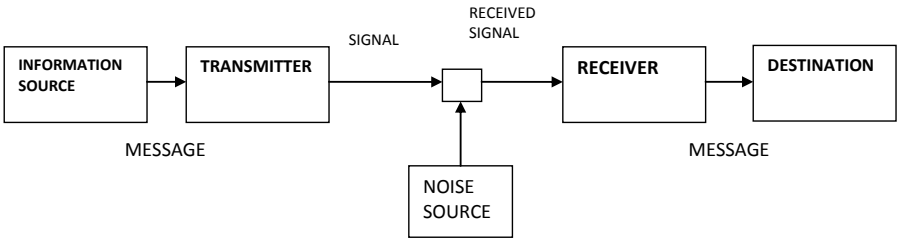


Figure 1.2: Schematic diagram of a general communication system [60].

well. The input signals are isolated nonsense syllables, which are received and decoded by the brain (the receiver). The transmitter is the speech production system (containing the vocal tract, mouth, lips and other articulators). Understanding how something breaks frequently gives important information about how it works. With this approach in mind, we attempt to disrupt the speech communication channel with different kinds of degradation: time truncation, filtering and noise-addition. This is the innovative three-dimensional deep search (3DDS) method, described in detail in [34]. Using the 3DDS, our group has isolated the precise perceptual cues used by normal listeners to perceive natural CV syllables. My research is concerned with understanding how different kinds of hearing losses degrade the use of these cues in HI listeners.

1.3 Thesis outline

The goal of this thesis is to gain fundamental insight into HI speech perception so that more advanced speech enhancing algorithms can be developed to compensate for hearing loss. The working hypothesis is that HI listeners may have difficulty with noisy speech because they cannot hear certain sounds, for which the characteristic features are lost due to both external noise and

hearing loss. The distorted speech cues may reduce the HI listeners' binaural processing ability and make it difficult for them to follow conversations in noisy environments. To explore the hypothesis, the following tasks are addressed sequentially. Background and literature review for each of the following tasks is included in the respective chapter.

Chapter 2 discusses some of my analyzes on normal hearing data. Per-utterance analysis explains the exponential nature of the Articulation Index theory. It also puts a bound on errors made by normals, and we show that errors are essentially zero for well-articulated tokens. This quantifies the robustness of human speech recognition.

Chapter 3 is the crux of our research and reports how hearing-impaired listeners perform on nonsense syllable recognition task in speech-weighted noise. With ample data from 46 HI ears, we show that HI speech perception is consonant dependent. We argue that the difficulties experienced by hearing-impaired listeners in noise come down to only a few consonant confusions and no other speech test used today is robust enough to capture this consonant dependence.

Chapter 4 is a brief overview of cochlear dead regions (dead inner hair cells region in HI ears). We present a new method based on comodulation masking release to detect these regions.

Chapter 5 presents results of HI speech test with NAL-R amplification. We show that this amplification scheme does not uniformly yield better scores for all CVs. It degrades performance of certain CVs for a particular HI ear. We stress the need for a new compensation scheme that is based on the perceptual cues being missed by the ear.

Chapter 6 is a summary of the main conclusions from our research.

CHAPTER 2

NORMAL HEARING SPEECH PERCEPTION

This chapter provides insight into the causes of consonant errors when listening in low-noise levels. We present the results of a closed-set recognition task for 24 stop consonant-vowel (CV) sounds (6 C $\times$ 4 V), spoken by 18 talkers in speech-weighted noise. We analyze the stop consonant errors on a per-utterance basis in low-noise environments, defined here as -2 dB signal-to-noise ratio (SNR), and in quiet. This per-utterance analysis shows that the error is zero for most (62.8%) plosive utterances. The remaining utterances (37.2%) that have errors can be classified into three groups - the high error (HE) group (10.7%) which contains the errors because of poor articulation of the utterance by the talker, the low error (LE) group (15.8%) that has utterances with low-grade (less than 1/190) random errors and the medium error (ME) (10.7%) group containing the remaining utterances. Consonant /b/ has the highest number of HE utterances. We hypothesize that normal hearing (NH) speech perception scores depend on the audibility of the CV feature, and the error is essentially zero when this feature is above threshold. The *Articulation Index (AI)* model, on the other hand, predicts an exponential average phone error $e = e_{min}^{AI}$ (e_{min} is the minimum error under ideal conditions, AI=1), due to the distribution of several utterances of a CV having different thresholds for their perceptual feature, the average of which approximates an exponential.

2.1 Introduction

Articulation Index theory was created by Harvey Fletcher in 1921 at Western Electric Research Labs to characterize the information-bearing frequency dependent regions of speech [2]. He modeled nonsense syllable recognition in terms of the average nonsense phone recognition score and found that the model did a good job of characterizing the raw data from a large number of measurements [18]. On the basis of psychoacoustic experiments with a large number of trials, Fletcher and his colleagues found that *the average nonsense phone articulation for CVC syllables,* defined as

$$s \equiv (2c + v)/3, \tag{2.1}$$

well represents nonsense CVC syllable recognition, defined as

$$S_3 \equiv c^2 v \approx s^3, \tag{2.2}$$

where c and v are consonant and vowel articulation scores, defined as the probability of maximum entropy (nonsense) sounds, respectively, as a function of the signal-to-noise ratio (SNR). Likewise, CV and VC syllable scores are accurately modeled as $S_2 \equiv cv \approx s^2$. The details of *Fletcher's methods of recognition* are documented by Allen [1] and Allen [2].

2.1.1 The AI model of average speech errors

Following the success of the *average phone score model* (Eq. 2.2), Fletcher extended the analysis to account for the effects of filtering the speech into bands [18]. This method later became known as *articulation index theory*,

which later became the well-known ANSI standard for Articulation Index (AI).

In general, for the fullband speech divided into $K = 20$ bands,

$$e = e_1 e_2 ... e_K, \tag{2.3}$$

where $e = 1 - s$ is the average fullband error, s is the fullband articulation (Eq. 2.1), and $e_k = 1 - s_k$ is the error in the k^{th} band where s_k is the band articulation. Thus, the band errors are treated as independent. Though the value of $K = 20$ was chosen empirically, it was later shown that each of these 20 articulation bands corresponds to approximately 1 mm along the basilar membrane [18] and the articulation density per critical band is a constant [1, 2].

The multiband product rule (Eq. 2.3) is also called *the additive law of frequency integration* and is the foundation of the ANSI standard for Speech Intelligibility Index (SII). This rule not only works for the average nonsense syllable score, but also fits the individual scores for 8 out of 16 (50%) Miller-Nicely consonants, namely /p,k,f,ʃ,b,d,g,z,m,n/, clearly as shown in the studies of [4] and [32].

Based on this assumption of independent articulation bands, [20] further developed a method for calculation of AI based on the intensity of the long-term average speech and noise. They extended Fletcher's original formulation by providing a formula for relating the band error (e_k) to the normalized signal-to-noise ratio in that band (SNR_k)

$$e_k = e_{min}^{SNR_k/K}, \tag{2.4}$$

where e_{min} is defined as the minimum error with ideal conditions (when AI

= 1). The total error is then

$$e = e_{min}^{AI}, \tag{2.5}$$

where $AI = \overline{AI}_k = \frac{1}{K}\sum_{k=1}^{K} AI_k$. Thus the total articulation is the sum of the band articulation over the K bands. The details of computing the normalized signal-to-noise ratio in each band (SNR_k) are described in [20].

From Eqs. 2.4 and 2.5, we see that e_{min} is an important parameter in the AI model. Quoting from the Acknowledgments section of the paper [4]: "The inspiration for this work started with a question by David Nahamoo which I could not answer: What is the meaning of e_{min}? " While the 2005 paper provides useful insight into how speech is processed by the auditory system and quantifies the nature of nonsense syllable confusions, some critical questions still remain unanswered, namely:

1. What is the meaning of e_{min}?

2. What is the nature of speech errors humans make in quiet?

3. Why does the Articulation Index model fit so well for certain specific classes of nonsense syllables?

Several variations of the AI model are extensively used to predict hearing-impaired speech perception as well [15, 49, 24], to characterize SNR-loss [27] and for hearing-aid fitting [56]. Generally speaking, while it is widely recognized that AI theory does an excellent job of characterizing the mean score, little is known as to why and how it works. Even more important, there is presently no theory that might predict consonant confusions [33]. This study is intended to partially close these gaps in our present understanding.

2.1.2 Capacity and error

[3] likened the AI model to Shannon's [60] concept of *channel capacity* and suggested this similarity is a fundamental information-theoretical basis for the empirical success of the AI theory. According to the *channel-capacity theorem*, the error goes to zero (there is complete transmission of information) while operating below capacity. If this is true for the human speech communication channel as well, then why is it that human listeners still make errors when there is no substantial additive noise? What is the nature of the error as a function of SNR? Namely is it zero below some threshold, as suggested by the channel capacity theorem, or does it go exponentially to e_{min} as given by Eq. 2.5? We will empirically address these two related questions by analyzing a database that consists of 25 normal hearing subjects responding to nonsense Miller-Nicely [41] CV syllables, at various levels of speech-weighted noise. The present analysis will be on a per-utterance basis rather than looking at the average score over each consonant, or across all consonants.

2.1.3 Aim of the study and approach

This chapter analyzes the six stop consonants, /p, t, k, b, d, g/. In the past, there have been several studies on perception of stop consonants. Benkí [5] studied the effects of place of articulation and F_1 transition on CV and VCV stimuli generated using the Klatt synthesizer [30]. Other important studies on synthetic plosives include Likser [35], Sumerfield and Haggard [64], Massaro and Oden [39] etc. Perceptual invariance in stop consonants has been extensively analyzed in classic studies, such as Blumstein and Stevens [6] and Stevens and Blumstein [63]. These studies necessitated the use of

synthetic syllables, so that one could control for the various acoustic cues and design desired templates.

However, we wish to answer questions like: Why are /pa/'s from some of the talkers confused with /ta/, while others are rarely confused? or Why are some speech sounds of the same consonant more robust than others? To answer these questions and to understand the robustness of human speech perception, one needs to account for the variability of natural speech, due to the talker, accent, masking noise, etc. Synthetic speech is typically of low quality and does not have this natural variance. We wish to retain the natural variability of the speech and the features, produced naturally by the vocal apparatus, to be able to answer the questions we are interested in. Hence, it is critical to use natural speech stimuli for our psychoacoustic experiments. We believe that we may answer these questions by studying natural speech randomly drawn from large databases of non-word syllables, so that we may gain fundamental insight about normal hearing (NH) speech perception.

Studies by [12] and [25] explored the same six plosives as used in our study, in the presence of three vowels (/a/,/i/,/u/). Both studies used natural speech produced by 2 male and 2 female talkers. While, they classified the errors in terms of the gender of the talker, they did not discuss the differences between the two talkers of the same gender. Both studies reported that most of the syllables had 100% correct responses in the absence of noise.

In the present study, we analyze the errors, made in low-noise condition in great detail, teasing apart the errors on a per-utterance basis. We will come to strong conclusions about NH speech perception, using a database having a large number of talkers (18) and listeners (25). We quantify the errors in terms of the percentage error and show that most utterances have "zero error." Per-utterance analysis provides a graphical way of categorizing the

utterances into the "zero-error" (ZE) group and the "non-zero error" (NZE) group. Next, we look closely at the type of errors made in the second group. Using percentage error, we classify the utterances into three groups: (1) a low error (LE) group (which essentially puts these utterances into the ZE group), (2) a high error (HE) group (due to talker misarticulation) and (3) a medium error (ME) group (which may be a combination of several factors like random errors, listener biases, misarticulated utterances and the effects of noise).

On the basis of our current analysis, we provide evidence that the average error (e of Eq. 2.3) is the average over a bimodal distribution of ZE and NZE utterances. For individual utterances, the error is *not* an exponential function of SNR, as given by Eq. 2.5. Rather, the score depends on the audibility of its primary cue [34] and the error is zero (except for low-grade random errors) as long as the perceptual feature of the utterance is above the noise floor. We shall see that different utterances have a natural distribution of thresholds for their perceptual features, and it is this distribution over a large number of utterances that makes the average error an exponential function of the signal-to-noise ratio. This view is consistent with the discussion provided by French and Steinberg [20].

The database used for the current study is the same as that used by Phatak and Allen [50] a.k.a. PA07. The aim of PA07 was to characterize consonant and vowel confusions in speech-weighted noise. For this purpose, they selected low error utterances (ones with less than 20% error in quiet) and the top ten "high-performing" (HP) listeners . It was necessary to remove these high error utterances and poor performing listeners, to study the impact of noise on consonant recognition. In the present study, we use the same database and corpus but revisit the errors for all the utterances, with par-

ticular emphasis on the high error utterances that were discarded by Phatak and Allen. The aim of this study is to categorize *all* the errors and to characterize every utterance. Hence the data to be analyzed includes *all* the utterances and *all* the listeners of PA07.

2.2 Methods

2.2.1 Stimuli

The experimental corpus is the same as reported by [50] and is called MN64 (MN because it is based on the classic Miller and Nicely experiment [41], and 64 because the database has 16C $\times$ 4V). A subset of isolated CV sounds from the LDC2005S22 corpus [19], recorded by the Linguistic Data Consortium (University of Pennsylvania), was used as the speech database. This subset had 18 talkers speaking CVs composed of one of the 16 Miller-Nicely [41] consonants (/p/, /t/, /k/, /f/, /θ/, /s/, /ʃ/, /b/, /d/, /g/, /v/, /ð/, /z/, /ʒ/, /m/, /n/), followed by one of the four vowels (/ɑ/,/ɛ/,/ɪ/,/æ/). Due to lack of IPA symbols in MATLAB, these vowels are referred to, in the figures and tables, as /a/, /e/, /I/ and /@/ respectively. Their error curves in all the relevant figures in this chapter have colors red, green, blue and yellow respectively. The vowels were chosen to have formant frequencies close to each other, with the goal of making them more confusable. All talkers were native speakers of English. Ten talkers spoke all 64 CVs, while each of the remaining eight talkers spoke different subsets of 32 CVs, such that each CV in MN64 was spoken by 14 talkers. Thus the experiment had 56 (14 talkers $\times$ 4 vowels) utterances of each CV. For the current analysis, we analyze a subset of the experimental data, i.e. only the stop consonants (/p/, /t/, /k/,

/b/, /d/, /g/).

For the experiment, the wideband noise rms level was adjusted according to the rms level of the CV sound to be presented, to achieve the required SNR. While calculating the rms level of a CV utterance, the samples -40 dB below the largest sample were removed [50].

2.2.2 Listeners

In total, there were 25 L1=English listeners (12 M and 13 F), having no known history of hearing disorder or impairment, and self-reported to have normal hearing. Fourteen listeners completed all the 42 sessions of the experiment which contained 5376 tokens. These 14 were the listeners reported in the PA07 study. Out of the remaining 11 listeners, 3 repeated a session resulting in $5376 + 128 = 5504$ tokens. The remaining 8 listeners completed fewer than 42 test sessions (the minimum being 4 and the maximum being 23). The average number of trials per CV per SNR is about 1060. Since there are 56 CV utterances, about 18-19 listeners (1060/56) heard a particular utterance at each SNR, on average.

2.2.3 Testing paradigm

The full test procedures are as described in [50]. All listeners were asked to identify the consonant and the vowel in the presented CV syllable by selecting one of 64 software buttons on a computer screen, each labeled with an individual CV. The 64 buttons were arranged in a 16×4 grid. The isolated CVs were played at six signal-to-noise ratios of [-22, -20, -16, -10, -2] dB and in quiet, in speech-weighted noise, the spectrum of which is described in [50]. A 'noise only' button was allowed if the participant heard only noise without

hearing any speech sound. These responses were treated as chance errors and distributed uniformly among the 16 possible responses, while scoring for the consonant (chance = 1/16). Of the total number of trials of the stop consonants across the 25 listeners, the percentage of 'noise only' responses was 0.03, 0.03, 0.15, 4.4, 29.2 and 46.8 respectively for quiet, [-2, -10, -16, -20, -22] dB SNR. Thus, this button was rarely used in the low-noise environment. Listeners heard the stimuli via headphones (Senheiser, HD-265). The listener was allowed to replay the CV sound as many times as desired before entering the response. Repeating the sound helped to improve the scores by eliminating the unlikely choices in the large 64-choice closed-set task and by allowing the listener to recover from the distractions during the long experiment. After the response button was clicked, the next sound was played following a short pause. Each presentation of CV sound was randomized over consonants, vowels, talkers, and SNRs. All 5376 presentations (16C $\times$ 4V $\times$14 talkers $\times$ 6 SNRs) were randomized and split into 42 tests, each with 128 sounds. Each listener was trained on the stimulus set using one or two practice tests with randomly selected sounds, presented in quiet, with visual feedback on the correct choice.

Each utterance was presented only once to a listener at each SNR, excluding the practice sessions. Since 14 listeners completed the task, the number of times a particular utterance was presented at a given SNR is at least 14. On an average, about 18-19 listeners heard a particular utterance (since the presentations are totally randomized, every listener who did not complete the task missed hearing a random set of utterances).

There is no substantial difference between scores in quiet and at -2 dB SNR for the stop consonants. Hence, we pool the data from these two conditions and define SNR $\geq$ -2 as the *low-noise environment*. Hence, the number of

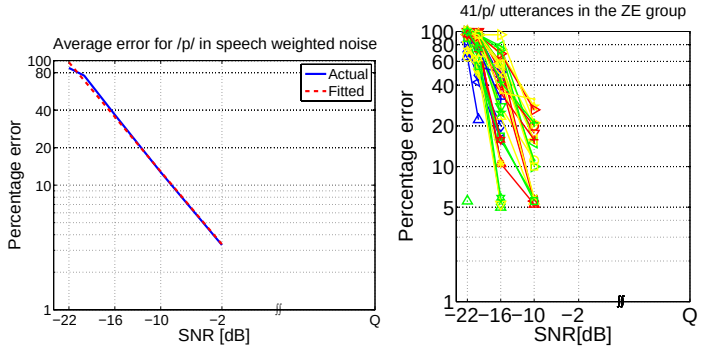
times (N) a particular utterance was heard in the low-noise environment is utterance dependent and is ≈ 38 on average. The actual value of N for each utterance is tabulated along with the utterance errors in the results section.

2.3 Per-utterance analysis of the raw data

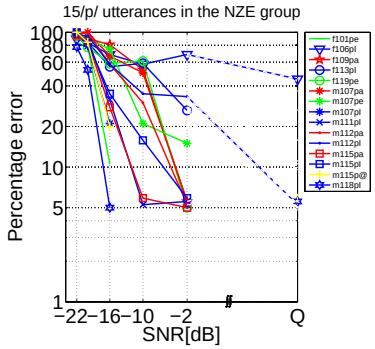
2.3.1 The AI model predictions

According to *the articulation index theory* and as shown in Fig. 2.1(a) for /p/, the *average* sound articulation error is given by Eq. 2.5. Hence, the average error is an exponential function of the AI. For speech-weighted noise (MN64), the AI is proportional to the SNR [4]. Hence, according to the AI theory, for speech-weighted noise, average error is an exponential function of the wideband SNR. This average is over talkers and listeners. In the previous study on MN64 by [50], the authors showed that the AI model fits the average error for three subsets of consonants: a low-scoring set C1: (/f/, /θ/, /v/, /ð/, /b/, /m/), a high-scoring set C2: (/t/, /s/, /z/, /ʃ/, /ʒ/) and set C3: (/n/, /p/, /g/, /k/, /d/) with intermediate scores. The respective values for e_{min} for these three groups are 0.01, 2×10^{-5} and 3×10^{-5}.

The AI model also works for 12 out of 16 consonants using a refined expression for AI, as shown in the same study. Figure 2.1(a) shows the average /p/ error as a function of SNR [dB] on a log scale in speech-weighted noise (SWN). The AI model precisely fits this via linear regression. Since, we do not know the actual SNR in the quiet condition, we cannot extend the total error of the graph to Q.



(a) Average /p/ error vs. SNR (b) Zero error (ZE) utterances



(c) Non-zero error (NZE) utterances

Figure 2.1: (a) Log error vs. SNR for /p/. The solid line represents the actual error, while the dashed line is the fitted regression line, bound by maximum error of 100%. This error curve for /p/ is the average of 56 different /p/ utterances, 14 for each of the 4 vowels. The error of panel (a) represents the total error e for /p/ as described in Eq. 2.3, and is consistent with Eq. 2.5 with $e_{min} \approx 0.03$ (3%). The two curves are almost identical (correlation coefficient = 0.99). Thus, the average log-error is linear [4]. Since we do not know the actual SNR of quiet, the line cannot be extended beyond -2 dB SNR and the discontinuity between -2 dB SNR and quiet (arbitrarily marked at 18 dB SNR) is marked by $\iint$ to the right of -2 dB SNR on the abssica. (b) The 41 zero error (ZE) utterances, defined as the sounds having no error at -2 dB SNR and quiet. (c) The non-zero error (NZE) utterances, ones that have errors at either -2 dB SNR and/or in quiet (i.e., low-noise). Quiet is indicated at 18 dB SNR on the abscissa and the error in quiet is joined to -2 dB points by dashed lines.

2.3.2 Individual utterance errors

The picture is totally different if we look at the individual utterances. Figures 2.1(b) and 2.1(c) show a breakdown of the 56 different utterances for the syllable /p/. Most of the utterances (41 out of 56) have zero error (ZE) at -2 dB SNR and above, while the other 15 are non-zero error (NZE).

2.3.3 Conflicting cues and priming

As demonstrated by [33], natural speech sounds, in particular the stop consonants, often contain *conflicting cues*, defined as significant energy at frequency regions representative of non-target plosives, characteristic of confusable sounds. There are frequent examples in the LDC corpus, where the talker poorly pronounces the target utterance. As a result, for these utterances the main perceptual feature (denoted *event*) is not robust and a conflicting cue dominates, even at low levels of noise. Such ambiguous sounds lie on confusion boundaries.

In addition to having conflicting cues, several plosive utterances have timing issues, where the cues are much closer to the start of the vowel, making the utterances susceptible to confusion with their voiced counterparts. Because of these misarticulations, a sound may not be robust, making it inherently ambiguous (or primable), even at low-noise levels. In such situations, normal hearing (NH) subjects must guess among a small group of confusable sounds, resulting in a low-entropy error group [4]. High error and low entropy are characteristic of such utterances. As shown in several examples in the following sections, these sounds may be identified on the basis of their error in low-noise conditions and error entropy, and can be explained by their AI-gram (Appendix A of [34]). Knowing the speech cues of a CV, we can

calculate the thresholds of the primary and conflicting cues of an utterance from the AI-gram, and can reliably predict the SNR at which the utterance will be at a confusion boundary, thus perceptually ambiguous. Such is the power of precise knowledge of speech cues.

2.3.4 Analysis methods

We next explain our terminology used in this thesis. We also discuss the criteria and rationale behind classifying the errors into groups. Figure 2.2 is such a grouping scheme for the consonant /p/. Table 2.1 gives the details for the NZE sounds.

1. There are 56 utterances for each consonant. The *low-noise environment* is defined as the SNR condition above -10 dB, i.e. -2 dB SNR and quiet. There is no substantial difference between these two conditions, hence the data are averaged across -2 dB SNR and quiet, to increase the utterance sample size N, given in Table 2.1.

2. N is the total number of presentations of each utterance in the low-noise environment. Since, at a given SNR each listener hears the sound only once, N is equal to the number of subjects who heard the CV at -2 dB SNR and in quiet. The average value of N is 38. The error distribution (e.g., mean and variance) mainly depends on N and the entropy (size) of the error group.

3. On the basis of errors made in the low-noise environment, the 56 utterances are divided into two groups: the zero error group (ZE) which contains utterances that have zero errors in the low-noise environment and the non-zero error (NZE) group, having at least one error in the low-noise environment.

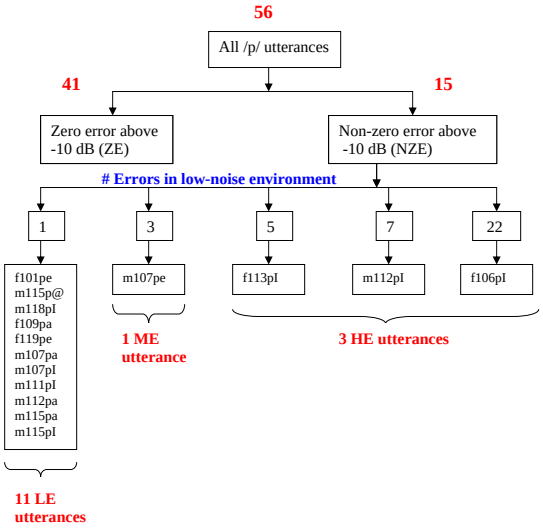


Figure 2.2: Error distribution of 56 /p/ utterances in the low-noise environment: The total number of utterances as marked above the topmost block is 56 (14 utterances each with 1 of the 4 vowels). The zero-error (ZE) group is on the left and contains 41 out of the 56 utterances as marked above the block. The number above a block gives the size of the group, i.e. number of utterances out of 56 that belong to that group. Out of the remaining 15 $(56 - 41)$ utterances, the next level shows the absolute number of errors made in the low-noise environment. As in the figure, 11 utterances have only 1 error (out of 38 trials on average), thus forming the low error (LE) group. Four utterances (m107pe, f113pI, m112pI and f106pI) have 3, 5, 7 and 22 errors respectively. The first two belong to the medium error group (ME), and the last two have greater than 12% error and therefore belong to the high error (HE) group.

Table 2.1: Percentage error, N and SNR_{90} values for in-error utterances of /p/. The table is divided into three groups with horizontal lines. The top 11 utterances have exactly 1 error ($< 3\%$). We interpret these errors as random. The last three utterances (f113pI, m112pI and f106pI), having more than 12% error, belong to the high error (HE) group. Utterance m107pe is a lone member of the medium error (ME) group. The SNR_{90} (the SNR at which the score drops from 100% to 90%) is highly correlated with the event threshold (Fig. 6a from [57]) and can be taken as an objective measure of the robustness of the sound. As seen from the tabulated values, ME and HE utterances have high ($\geq$ -2 dB) SNR_{90} thresholds. Thus, they are easily confusable, even in low-noise environment. In particular, f106pI has more than 50% error even in quiet, so its SNR_{90} value is ∞. LE utterances have low values for SNR_{90} ($<$ -2 dB)and hence are *robust* and are classified as being in the robust zero error (RZE) group.

utterance	P_e [%]	N	SNR_{90}
f101pe	2.70	37	-16
m115p@	2.78	36	-14
m118pI	2.78	36	-16
f109pa	2.78	36	-3
f119pe	2.56	39	-3
m107pa	2.70	37	-3
m107pI	2.70	37	-12
m111pI	2.86	35	-12
m112pa	2.56	39	-4
m115pa	2.70	37	-12
m115pI	2.70	37	-5
m107pe	7.69	39	5
f113pI	13.89	36	10
m112pI	18.92	37	15
f106pI	56.41	39	∞

4. P_e is the error, in [%] units, for an utterance in the low-noise environment.

5. $\mathcal{H}_N$ is the *normalized entropy* of an utterance, normalized by the maximum possible entropy. The intuition for using this measure will be explained with an example in the next section.

6. Based on our error analysis, the utterances in the NZE group are divided into three groups: low error (LE) group, medium error (ME) group and high error (HE) group.

7. The LE group contains utterances with $\leq 3\%$ error in low-noise environment. This corresponds to a single error in N (e.g. 38) trials. We hypothesize these errors are uncorrelated across listeners and vowels, hence are random. Later, we shall show that this is not exactly true; the LE rate depends slightly on the difficulty of the task.

8. The HE group contains utterances with $P_e \geq 12\%$. We term these errors as *true*. We anticipate that these errors are because the utterances are ambiguous due to poor articulation by the talker. We show that these *ambiguous* utterances form a low-entropy confusion group.

9. The ME group contains the remaining utterances having $3\% < P_e < 12\%$. It is difficult to come to a precise conclusion about these utterances without either more data, or alternatively, a more extensive analysis than provided here. Since very few utterances fall into this group, we will not analyze them further. However, we conjecture that these errors are because of a combination of many factors, such as random errors, listener biases, misarticulated utterances and the effects of noise.

10. The ZE and the LE group together define the *robust zero error* (RZE) group. The utterances in this group are called "robust" because they either have no errors or random errors that are inherent in any experiment.

11. The utterances in the HE group are called "ambiguous" because these are due to poor articulation by the talker, are easily heard by most listeners and confusable within a low entropy (small) group. This group of sounds is easily *primed.* The term *priming* is used as a test of the natural ambiguity of a phone, as discussed in the text.

Let $|G|$ denote the cardinality of group G. Thus,

$$56 = |ZE| + |NZE|,$$

$$|NZE| = |LE| + |ME| + |HE|,$$

Number of *robust sounds* $= |RZE| = |ZE| + |LE|$.
Number of *ambiguous sounds* $= |HE|$.

2.4 Results

2.4.1 Error groups for the unvoiced plosives

This section analyzes the three unvoiced plosives,/p/,/t/,/k/, on an utterance-by-utterance basis using the methods developed in Section 3.4. In Section 4.2, we analyze the unvoiced plosives /b/,/d/,/g/.

2.4.1.1 Error analysis for /p/

As shown in Fig. 2.1(b), out of a total of 56 /p/ utterances, 41 have zero error in the low-noise environment (ZE group). The remaining 15 utterances (NZE group) are characterized in Table 2.1, where we also tabulate N. We analyze this NZE set based on the number of errors made in the low-noise environment (-2 dB SNR and quiet). The topmost box (the top level in the figure) gives the total number of utterances, which is 56. For the first cut, we classify these utterances into two groups:

1. The zero error (ZE) group: These utterances have 100% score in the low-noise environment. Forty-one /p/ utterances (73%) fall into this group. This is marked on the leftmost box on the second level in Fig. 2.2.

2. The non-zero error (NZE) group: These utterances have an error in the low-noise environment. Fifteen /p/ utterances (27%) fall into this group, as indicated above the rightmost box on the second level in the figure.

Next, we break down the NZE utterances in terms of the number of errors made in the low-noise environment. As shown in Fig. 2.2, 11 utterances have 1 error while the remaining 4 utterances have 3, 5, 7 and 22 errors respectively.

Thus, we see that no listener makes any errors on 41 of the utterances. Of the remaining 15 utterances, 11 have one error (less than 3% error), across ≈ 38 listener trials. We call these single error utterances "random errors". These utterances are well-articulated and the errors must be random, since most listeners (i.e. $N-1$ out of N) get them right. If the experimental

trials were repeated we would expect this list of sounds to totally change, as they reflect the random error rate. As shown later, we estimate that for /p/ a listener makes a random error once every 190 trials or so, on average. Possible causes of these errors are lack of attention, wrong button clicked, etc. Such "cosmic ray" errors are expected and impossible (or at least difficult) to control. LE utterances are not inherently ambiguous (cannot be primed), rather they have random low-grade errors and we view them as belonging to the ZE group. Priming is defined as mentally selecting the consonant heard by making a conscious choice between several possibilities having neighboring scores [57]. This ZE $\cup$ LE group, defined as the robust zero error (RZE) group, contain 52 $(41 + 11)$ out of 56 /p/ utterances (92.8%).

The ME group contains a lone utterance m107pe which has 3 errors out of 39. These 3 errors are all at -2 dB SNR and had confusions /f,g,ʒ/. We presently have no clear intuition about the underlying nature of these errors. More data will be required to resolve the nature of the ME group.

The HE group for consonant /p/ contains the three utterances f113pI, m112pI and f106pI. The confusions for f113pI were /b,k,n,t,y/, for m112pI were /g,g,k,k,k,ð,ʒ/ while all the errors for f106pI (22 out of 39) were attributed to /t/. It is clear that f106pI is clearly ambiguous and at a /p-t/ confusion boundary. However the other two sounds, though high in error, are not consistent in their confusions across listeners.

A useful tool to characterize such confusions is the normalized entropy $\mathcal{H}_N$, defined as the entropy of the syllable divided by the maximum possible entropy for the given percentage error. For example, f101pe has one 1 error out of 37 presentations, so the error is 1/37 or 2.7%. The entropy $\mathcal{H}$ is

$$\mathcal{H} = -\left[\frac{36}{37} log_2 \left(\frac{36}{37}\right) + \frac{1}{37} log_2 \left(\frac{1}{37}\right)\right] = 0.179.$$

The maximum entropy for the given error would correspond to the error (2.7%) being uniformly distributed over the 15 bins, corresponding to all consonants except /p/. Hence, maximum entropy ($\mathcal{H}_M$) for this utterance is

$$\mathcal{H}_M = -\left[\frac{36}{37} log_2\left(\frac{36}{37}\right) + \frac{15}{15*37} log_2\left(\frac{1}{15*37}\right)\right] = 0.285.$$

Thus, the normalized entropy ($\mathcal{H}_N$) is $0.179/0.285 = 0.63$. $\mathcal{H}_N$ is a measure of randomness of the error. Had the error been totally random, $\mathcal{H}_N$ would be at its maximum value of 1. $\mathcal{H}_N$ values for f113pI, m112pI and f106pI are 0.80, 0.73 and 0.31 respectively.

Utterances with high errors and low normalized entropy are expected due to a talker misarticulation, which is heard by multiple listeners as confusable within a small confusion group. For example, the reason why most listeners (22 out of 39 trials) reported /t/ when f106pI was presented can be easily explained by looking at the AI-gram of the utterance. As seen from the articulation-index gram (AI-gram) at 12 dB SNR of Fig. 2.3, this utterance has significant energy above 4 kHz (rectangular box region), which is a /t/ cue [57], rendering this utterance ambiguous, as either /p/ or /t/. Thus, 22 times out of 39, it is reported as a /t/. Listening to this utterance, one can easily prime for this as /p/ or /t/, but no other consonant.

Another way of classifying the utterances is to look at the distribution of the *event thresholds*. Every utterance of a syllable has a different threshold for its *event* (perceptual feature). This threshold is called the event threshold (SNR_e). As demonstrated by [57], the event threshold is highly correlated with SNR_{90}, defined as the SNR at which the score drops from 100% to 90%. For example, f101pe is a robust utterance with its SNR_{90} at -16 dB, while m107pe is a weak utterance ($SNR_{90} \approx 5$ dB). As the noise increases, the

event is masked at SNR_e and the syllable becomes inaudible/confusable with the loss of its single primary feature. Correspondingly, the score falls from 90% to chance performance within a few decibels, below SNR_{90}. The SNR_{90} values of the NZE utterances are tabulated in Table 2.1. Some sounds never reach 90% score even in quiet, and for these SNR_{90} $=\infty$. The LE sounds have low values (< -2 dB) for SNR_{90}. Hence, these are robust in the low-noise environment and are classified as being in the RZE group. ME and HE utterances have high perceptual thresholds and, hence, these sounds have *true errors*, and are ambiguous.

If a sound were to have more than one event, the score would not drop rapidly. Thus, the very rapid drop in score below SNR_{90} proves that there is a single event, as is the case for /t/ [57].

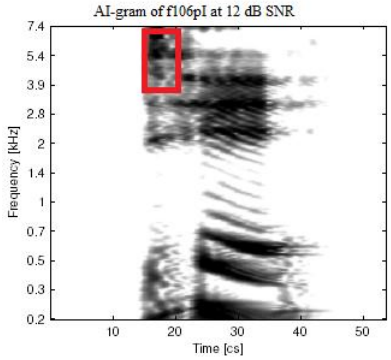


Figure 2.3: AI-gram of f106pI at 12 dB SNR. The conflicting cue is marked by a solid (red) box. This clearly shows a high frequency conflicting /t/ burst [33]. The utterance is primable as either /p/ or /t/. Correspondingly the error is 56%. The time axis is labeled in centiseconds [cs] because these units are highly relevant to speech perception. 1 cs = 10 ms.

2.4.1.2 Error analysis for /t/

Just as in the previous case of /p/, we analyze /t/ at the utterance level. As seen in Fig. 2.4 and Table 2.2, out of 56 /t/ utterances, 40 have zero error

in the low-noise environment (ZE group). Note that the SNR_{90} values for NZE utterances are marked as ∞ in the tables if the score does not reach 90% even in quiet. Also, for interpolation of the SNR_{90} values, quiet was marked at 18 dB SNR. Of the 16 NZE CVs, only two are ambiguous (belong to the HE group): m117te and f103te. Interestingly, all the errors in m117te (5/38) are /p/, resulting in the normalized entropy ($\mathcal{H}_N$) value of 0.52. Thus, this case is a natural complement to the case of f106pI, where all the /p/ errors were /t/. This again is predictable when one studies the AI-gram of m117te, which has a significant low-frequency energy, which is a conflicting cue region for /p/ [34, 33]. Utterance f103te (5 errors) is mostly confused with /d/ ($\mathcal{H}_N = 0.41$), because the utterance has a very short time-gap between the burst feature and the vowel, as is characteristic of voiced /d/ [34].

2.4.1.3 Error analysis for /k/

Out of 56 /k/ utterances (Fig. 2.5 and Table 2.3), 49 have zero error in the low-noise environment. Only 7 /k/ utterances are in error and only two of these, f101ka and f101kI, have high errors. Both are confusable with only one other sound: /g/. Talker f101 is a poor articulator for /k/. Figure 2.6 shows the AI-grams of these two sounds respectively. From the study by [34], the /ka/ cue is a mid-frequency burst around 2 kHz, articulated 5-7 cs before the vowel. On the other hand /ga/, the voiced counterpart of /ka/, has a mid-frequency burst, typically followed by a F2 transition just before the start of sonorance. As seen from the AI-grams of Fig. 2.6, f101ka has its burst cue just before the vowel start and does not have the characteristic 5-7 cs gap before the onset of the vowel, typical of a clearly articulated /ka/. Similarly, f101kI is atypical because unvoiced stops do not have bursts close

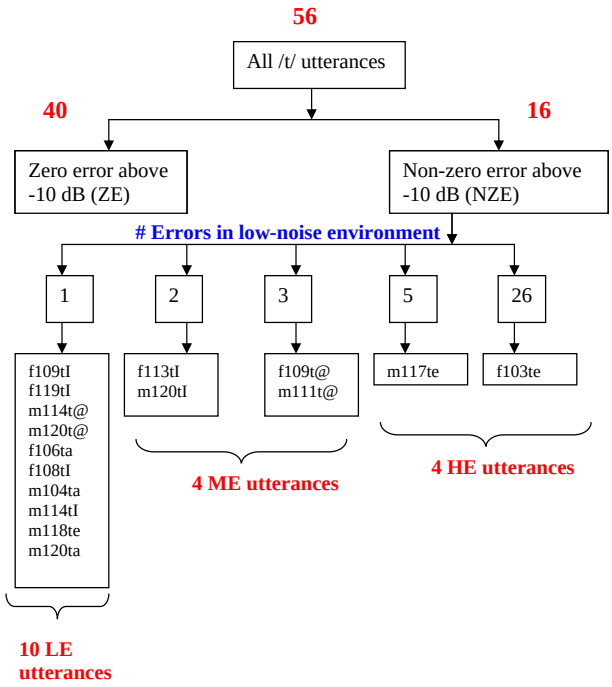


Figure 2.4: Error distribution of 56 /t/ utterances in the low-noise environment: The total number of utterances, as marked above the topmost block, is 56. The zero-error (ZE) group on the left contains 40 out of the 56 utterances (71.4%). Out of the remaining 16 $(56 - 40)$ utterances, the third level shows the absolute number of errors made in the low-noise environment. We see that 10 utterances have only 1 error (out of ≈ 38 trials on average across 25 listeners), defining the LE group. One sound (m117te) has 5 errors and another sound (f103te) has 26 errors. We place these sounds in the HE group. The remaining 4 $(16 - 10 - 2)$ are ME utterances, which we do not attempt to explain.

Table 2.2: Percentage error, N and SNR_{90} values for in-error utterances of /t/. Ten utterances in the topmost block with a single error (effectively less than 3% error) belong to the LE group, the next 4 in the middle block are ME utterances, while m117te and f103te are HE ambiguous utterances. The HE utterances have high SNR_{90} thresholds as seen in the table.

utterance	P_e [%]	N	SNR_{90}
f109tI	2.70	37	-22
f119tI	2.56	39	-14
m114t@	2.70	37	-11
m120t@	2.78	36	-21
f106ta	2.56	39	-11
f108tI	2.63	38	-22
m104ta	2.70	37	-10
m114tI	2.70	37	-17
m118te	2.63	38	-17
m120ta	2.70	37	-22
f113tI	5.26	38	-11
m120tI	5.13	39	-22
f109t@	7.89	38	-16
m111t@	8.11	37	-4
m117te	13.16	38	18
f103te	68.42	38	∞

to the vocalic region. Hence, these two sounds are confused with /g/. Vowel onset is marked by a solid (green) line, while the burst cue is boxed (red).

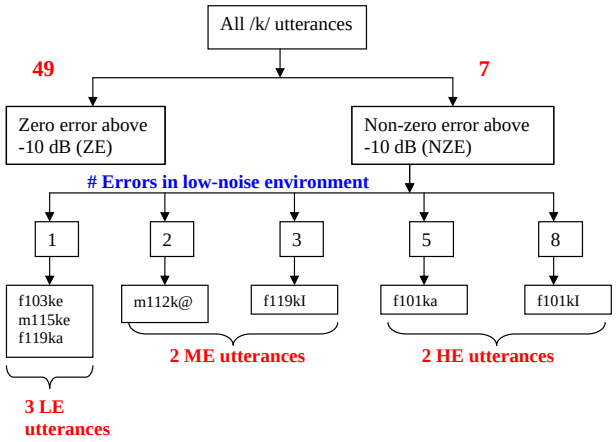


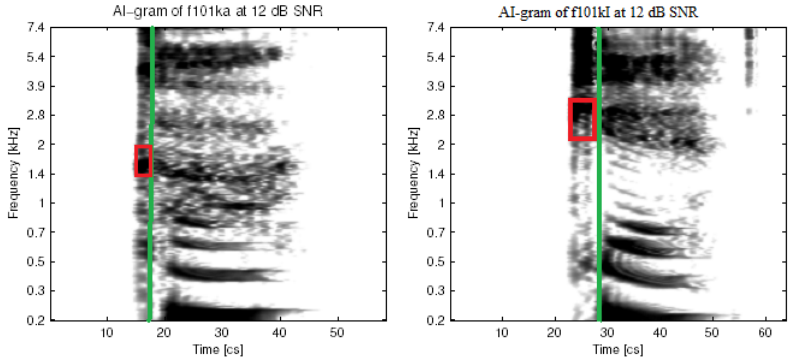
Figure 2.5: Error distribution of 56 /k/ utterances in the low-noise environment: /k/ is a robust sound with 49 out of 56 utterances in the ZE group. Only the two rightmost sounds (f101ka and f101KI) are ambiguous. Three utterances (f103ke, m115ke, f119ka) are LE while m112k@ and f119kI are in the ME group.

2.4.2 Error groups for the voiced plosives

As their unvoiced counterparts, the voiced plosives (/b/,/d/,/g/) also have utterances with different perceptual thresholds. Specifically, /b/ is a high error sound and forms a confusion group with the fricatives /v-f/, since the /b/ acoustic feature is not robust and is easily masked by noise. It is the lone plosive in the low-scoring set (C1) of the PA07 study [50]. One might perceive /b/ as having low *salience*. However, robust zero utterances with low SNR_{90} thresholds still exist but are rare (11 out of 56 utterances in this sample).

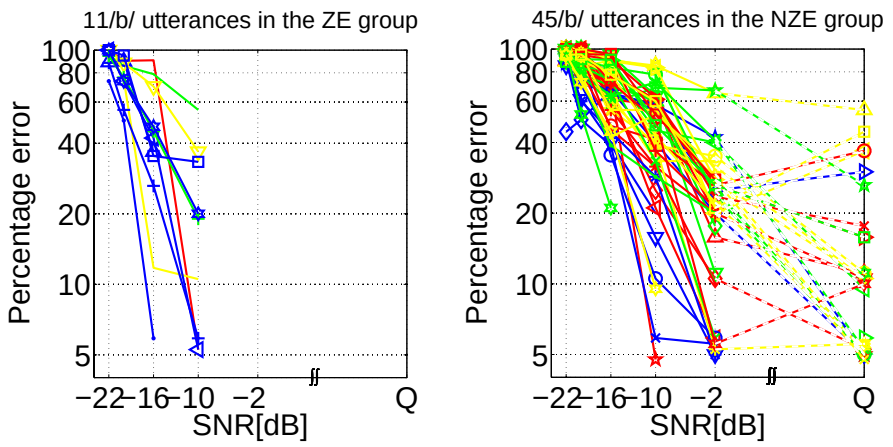
Table 2.3: Percentage error, N and SNR_{90} values for in-error utterances of /k/ in the low-noise environment. The NZE group is half that of /p/ and /t/. We interpret /k/ as having high *salience*, meaning it is easily articulated and easily identified (i.e. it is naturally robust). The top three utterances belong to the LE group, the next two to the ME group and the last two are HE utterances (with high SNR_{90} values).

utterance	P_e [%]	N	SNR_{90}
f103ke	2.56	39	-17
m115ke	2.56	39	-16
f119ka	2.63	38	-4
m112k@	5.13	39	-2
f119kI	7.89	38	-11
f101ka	13.89	36	18
f101kI	22.22	36	∞

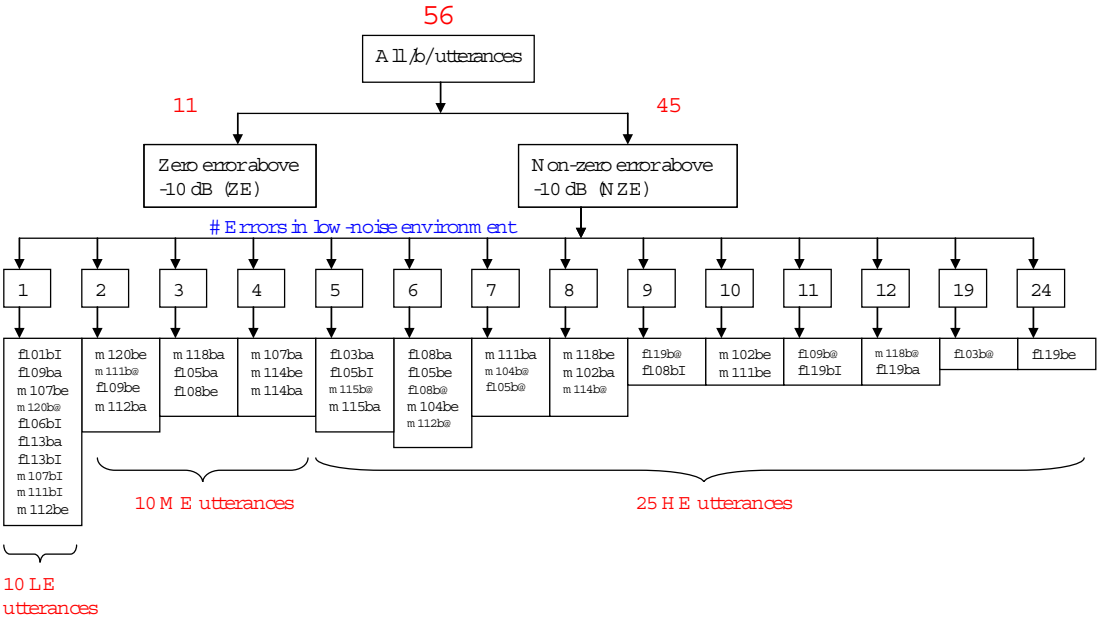


(a) AI-gram of f101ka. (b) AI-gram of f101kI.

Figure 2.6: AI-grams at 12 dB SNR. In both the AI-grams, the solid (red) box is the /k/ feature while the start of the vowel is marked by a solid (green) line. We see that the burst cue is very close to the beginning of the vowel, which is a characteristic of the /g/ feature [34], thereby explaining why these two /k/ utterances are highly confusable with /g/.



(a) Zero error utterances. (b) Non-zero error utterances.



(c) Error distribution of /b/ utterances.

Figure 2.7: The figure shows the distribution of errors of the 56 utterances of /b/. (a) Error vs. SNR plot of the 11 utterances that have no error in the low-noise environment. (b) Error vs. SNR plot of the remaining 45 utterances that have errors in the low-noise environment. Quiet is marked at 18 dB on the SNR scale and is joined to -2 dB SNR points via dashed lines. (c) Breaking down the errors in low-noise environment on the basis of the absolute number of errors made. Twenty-five out of 56 (44%) utterances have high errors (ambiguous utterances).

2.4.2.1 Error analysis for /b/

Eleven out of 56 /b/ utterances have zero error in the low-noise environment. The breakup of the utterances into the two main error groups, and the distribution of the errors in the second group (NZE), is shown in Fig. 2.7 and tabulated in Table 2.4.

Consonant /b/ is substantially different from the other 5 plosives used in the study, as it has much higher errors. In the low-noise environment, the error rates for /p/,/t/,/k/,/b/,/d/,/g/ are typically 1.8%, 2.3%, 0.8%, 11%, 2.2% and 0.7% respectively. Thus /b/ has low salience. We suspect that the high /b/ error is mainly a production error, the evidence for which is the 11 ZE utterances and fact that 13 out of 14 talkers of /b/ have the high-error utterances. Talker f101 has all its utterances in the ZE group. This proves that the listeners can do the task, since they make no errors for the talker (e.g. f101) clearly enunciates the consonant /b/. However, /b/ is difficult to articulate and is easily confusable (low salience). Unlike /t/ or /g/, it does not have a well-defined single feature that makes it noise-robust [34]. Instead of analyzing the three groups for /b/ as we have done for the other plosives, we try to analyze the errors on a listener basis (rather than on a per-utterance basis.)

We know that /b/ forms a confusion group with /f/ and /v/. These three consonants have high errors even in low-noise environments [41, 34]. As previously mentioned, we have assumed that the subjects in this experiment form a homogeneous group. While this is a reasonable assumption for the other low-error plosives, it seems to break down when the task becomes difficult, e.g. perception of /b/. A difficult test naturally categorizes the test-takers into performance groups. *When the going gets tough, only the*

Table 2.4: Percentage error, N and SNR_{90} values for in-error utterances of /b/. The horizontal line is the demarcation between 10 low error (LE) utterances (above) and the 10 medium error (ME) utterances (below). The entire right column of the table are the HE utterances (25 in total out of the 45 NZE utterances that have errors). Clearly, /b/ is a difficult sound compared to the other 5 plosives, since a majority of its utterances have high errors. Such high errors are likely to be due to production errors as evidenced by the fact that one talker (f101) has no error. This shows that the listeners can hear a well articulated /b/. For most HE sounds, /b/ is confused with /v/ and /f/. These utterances have high thresholds and most do not reach 90% score even in quiet.

utterance	P_e [%]	N	SNR_{90}	utterance	P_e [%]	N	SNR_{90}
f101bI	2.63	38	-6	f103ba	13.51	37	18
f109ba	2.63	38	-11	f105bI	12.50	40	5
m107be	2.70	37	-6	m115b@	13.51	37	12
m120b@	2.70	37	-10	m115ba	13.89	36	11
f106bI	2.70	37	-6	f108ba	15.38	39	18
f113ba	2.78	36	-4	f105be	15.38	39	13
f113bI	2.86	35	-10	f108b@	16.22	37	14
m107bI	2.50	40	-4	m104be	14.63	41	12
m111bI	2.86	35	-11	m112b@	15.38	39	13
m112be	2.86	35	-4	m111ba	20.59	34	∞
m120be	5.71	35	-16	m104b@	18.92	37	18
m111b@	5.41	37	-4	f105b@	17.95	39	12
f109be	5.56	36	0	m118be	21.05	38	∞
m112ba	5.00	40	-3	m102ba	21.05	38	∞
m118ba	7.89	38	-3	m114b@	21.05	38	18
f105ba	7.89	38	-2	f119b@	23.68	38	18
f108be	8.33	36	6	f108bI	23.08	39	15
m107ba	10.81	37	7	m102be	24.39	41	18
m114be	10.00	40	7	m111be	24.39	41	15
m114ba	10.53	38	8	f109b@	28.21	39	∞
				f119bI	27.50	40	∞
				m118b@	32.43	37	∞
				f119ba	31.58	38	∞
				f103b@	47.50	40	∞
				f119be	60.00	40	∞

tough can get going. In PA07, 4 low performance (LP) listeners, with scores less than 85% in quiet, were removed during analysis, and the top 10 high performance (HP) listeners were selected. Each of these 14 $(4 + 10)$ listeners completed the experiment (5376 tokens). Figure 2.8 shows the log-error versus SNR for consonant /b/ for these 14 listeners. The legend provides a two-letter listener ID. Of these, listener QN has the lowest error rate, except for quiet, suggesting a varying attention during the task. Subjects BH and LT have substantially higher error across SNR as compared to the average. In quiet, the listeners with greater than average error are BH, LT, SC, CO, QN and AN. Of these, BH, LT, QN and AN are the subjects (labeled on the figure) that were removed during the analysis in PA07. Thus, with the exception of QN, the poor performing listeners on average are also the poor listeners of /b/. The other 11 listeners who completed varying numbers of trials are not shown in the figure. However, these listeners also naturally break down into performance groups. For an easy task, there is much less of a difference between the low performance (LP) and high performance (HP) listeners. But these groups clearly stand out once the task becomes difficult. Thus, most errors on /b/ can be attributed to the LP subjects.

In summary, when the task is easy (i.e. for naturally low error utterances like /p/, /k/, /g/ etc., which have high salience), the main contributors to the error in low-noise environments (or to e_{min}) are a few ambiguous utterances. On the other hand, for high error and highly confusable consonants like /b/, /f/, /θ/ or /ð/ (low salience), there is a significant disparity among the listeners. Most utterances of high error syllables are erroneous primarily because of the LP subjects.

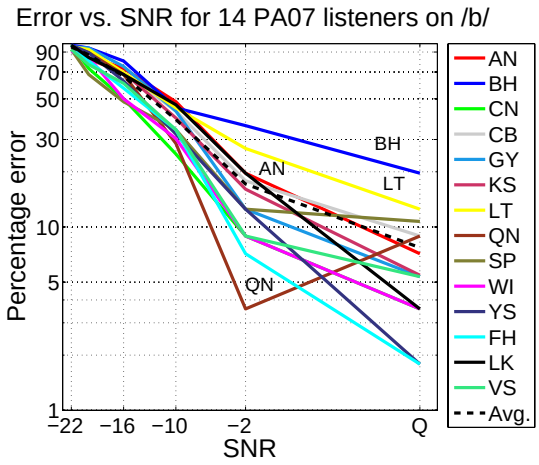


Figure 2.8: Log error vs. SNR for /b/ (average over 56 utterances) for the 14 listeners who completed the experiment (PA07). The grand average error over these listeners is shown by a dashed line. The legend indicates each listener with a two-letter ID. In quiet, the listeners with greater than average error are BH, LT, SP, CB, QN and AN. Of these, BH, LT, QN and AN are the four listeners removed from the analysis in PA07.

2.4.2.2 Error analysis for /d/

Out of 56 /d/ utterances, 27 have zero error in the low-noise environment. The distribution of errors is shown in Fig. 2.9 and tabulated in Table 2.5, which shows that /d/ has 12 utterances with random errors in the LE group, 13 in the ME group. Four utterances (m118d@, m102de, m115dI and m114d@) are characterized by high error and low entropy, and belong to the HE group. Out of these, m118d@ and m114d@ have timing errors and are confusable with the unvoiced counterpart /g/. m115dI has a conflicting cue of /b/ and is confused 7 out of 38 times with /b/ and once with /ð/. m102de is mainly confused with /ð/, perhaps because m102de is not articulated with sufficient "voicing." Errors on consonant /d/ are mainly production errors. Some listeners have difficulty phonotactically identifying the difference between /d/ and /ð/, possibly due to poor phonemic training, early in life.

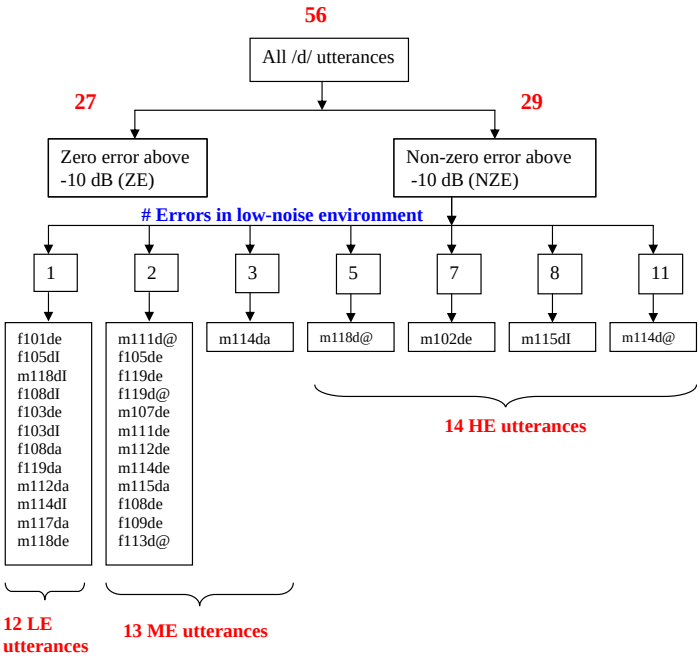


Figure 2.9: Error distribution of 56 /d/ utterances in the low-noise environment: The zero-error (ZE) group on the left-top contains 27 out of the 56 utterances (48%). Out of the remaining 29 (56−27) utterances, the third level shows the absolute number of errors made in the low-noise environment. Of these the 4 utterances to the right, m118d@, m102de, m115dI and m114d@ , having 5, 7, 8 and 11 errors respectively, belong to the HE group.

Table 2.5: Percentage error, N and SNR_{90} values for NZE utterances of /d/. The left 4 columns contain the 12 LE utterances. The horizontal line on the right 4 columns is the demarcation between the 13 medium error (ME) utterances (above), and the 4 high error (HE) utterances (below). The SNR_{90} values are well correlated with these three groups: LE sounds have low thresholds while HE sounds have high perceptual thresholds, even ∞ for sounds whose score does not reach 90% even in quiet.

utterance	P_e [%]	N	SNR_{90}	utterance	P_e [%]	N	SNR_{90}
f101de	2.63	38	-21	m111d@	5.41	37	-4
f105dI	2.78	36	-17	f105de	5.13	39	-11
m118dI	2.63	38	-13	f119de	5.26	38	-10
f108dI	2.78	36	-21	f119d@	5.00	40	-20
f103de	2.44	41	-13	m107de	5.41	37	-11
f103dI	2.86	35	-20	m111de	5.26	38	-15
f108da	2.44	41	-11	m112de	5.26	39	-17
f119da	2.78	36	-20	m114de	5.13	39	-4
m112da	2.56	39	-10	m115da	5.13	39	-3
m114dI	2.70	37	-17	f108de	5.13	39	-2
m117da	2.63	38	-10	f109de	5.41	37	-20
m118de	2.56	39	-10	f113d@	5.56	36	-10
				m114da	8.57	35	-3
				m118d@	13.89	36	∞
				m102de	17.95	39	12
				m115dI	21.05	38	13
				m114d@	27.50	40	∞

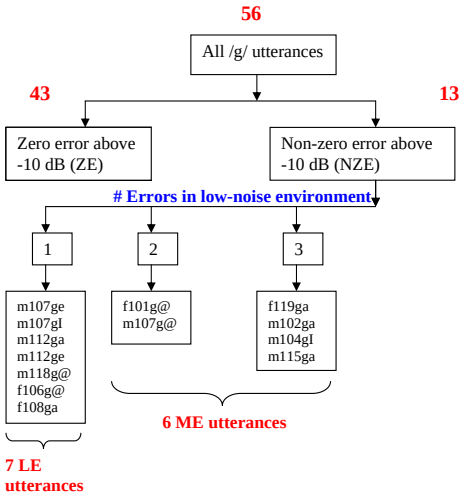


Figure 2.10: Error distribution of 56 /g/ utterances in the low-noise environment: 43 utterances are in the ZE group and 13 in the NZE group. Of these 13, 7 have a single (random) error while the other 6 have medium error. There are no HE sounds for /g/.

2.4.2.3 Error analysis for /g/

Out of 56 /g/ utterances, 43 have zero error in the low-noise environment. The distribution of errors is shown in Fig. 2.10 and the errors are tabulated in Table 2.6. /g/ is a robust (highly salient) sound and no utterance used in the PA07 experiment is misarticulated (i.e. no HE utterance), according to our criterion of $\geq 12\%$ in the low-noise environment.

Table 2.6: Percentage error, N and SNR_{90} values for NZE utterances of /g/. All the 56 /g/ utterances used in the experiment are well-articulated and have no high errors. The utterances in the left 4 columns form the LE group while the right 3 column utterances belong to the ME group. All NZE utterances have SNR_{90} threshold below -2 dB SNR.

utterance	P_e [%]	N	SNR_{90}	utterance	P_e [%]	N	SNR_{90}
m107ge	2.94	34	-11	f101g@	5.13	39	-13
m107gI	2.44	41	-12	m107g@	5.26	38	-13
m112ga	2.78	36	-11	f119ga	7.50	40	-7
m112ge	2.50	40	-12	m102ga	7.89	38	-3
m118g@	2.63	38	-13	m104gI	8.11	37	-10
f106g@	2.78	36	-5	m115ga	7.32	41	-3
f108ga	2.70	37	-3				

2.4.3 Error distribution across vowels

Out of the total 336 utterances (6 plosives × 56 utterances of each) in the experiment, 125 belong to the NZE group (15 for /p/+ 16 for /t/+ 7 for /k/+45 for /b/+ 29 for /d/+13 for /g/). Broken down by the vowel, they are 33, 34, 31 and 27 for (/ɑ/,/ɛ/,/ɪ/,/æ/) respectively. This gives an entropy of 1.99 bits. Thus the error distribution over the vowels is almost uniform (uniform distribution would imply a maximum 2 bit entropy). Thus, the consonant errors averaged across talkers are uncorrelated with the following vowel, in the low-noise environment.

2.5 Summary and discussion

Figure 2.11 summarizes the errors made by listeners on the six stop consonants. As one may see from the bar plot, /b/ has, by far, the largest number of utterances in the high error (HE) group. Hence, /b/ is a difficult sound (has low salience). The remaining five CVs have a few utterances that fall into the HE group and they represent a major component of e_{min}. By our definition, "robust utterances" = ZE + LE = RZE while "ambiguous" ut-

terances = HE utterances. The "ambiguous" utterances (HE) count (out of 56) is 3, 2, 2, 25, 4 and 0 for p, t, k, b, d and g respectively. On the other hand, most utterances are robust and have essentially zero error. The percentage of robust zero error (RZE) sounds is given by $\frac{|ZE+LE|}{56} \times 100$. This percentage for /p,t,k,b,d,g/ is 92.8%, 89.3%, 92.9%, 37.5%, 73.2% and 89.3% respectively.

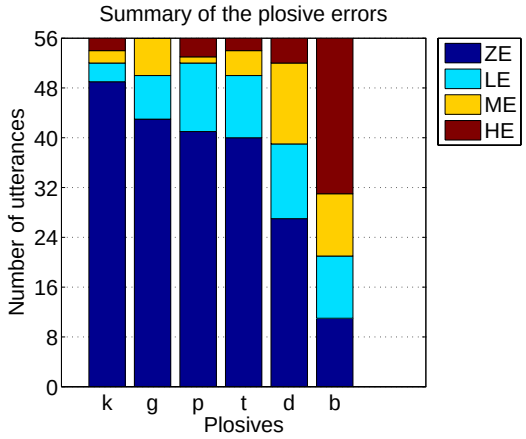


Figure 2.11: Stacked bar-plot gives the errors made by the 6 plosives in speech-weighted noise in the low-noise environment. The abscissa shows the 6 consonants, arranged in the order of decreasing number of utterances in the ZE group (the order of decreasing salience). The ordinate indicates the number of utterances of the consonant that falls into the ZE, LE, ME and HE groups respectively. The total is always 56. ZE is the zero-error group which contains utterances that all listeners had correct scores on, at -2 dB SNR and in quiet. LE is the low error group, having low-grade random error. ME is the medium error group with utterances having 3-12% error. HE group utterances have errors greater than 12% and are primarily due to production errors. These are always ambiguous/primable utterances with high errors and low entropy. ZE and LE groups together form the robust zero error (RZE) group.

Averaged across the 6 stop consonants, the percentage of utterances in the ZE, LE, ME and HE group is 62.8%, 15.8%, 10.7% and 10.7% respectively.

2.5.1 Theoretical considerations

Finally, we estimate the average number of trials needed by a listener before he/she makes a low-level random error. We assume that all 25 listeners are homogeneous (this is of course not true, since some listeners are significantly poorer than others, as demonstrated by [50]). Given that /p/ has a naturally low error ($|NZE| = 15$), it is reasonable to consider listeners as uniform. In total, 2121 tokens of 56 /p/ utterances were presented in the low-noise environment (1059 at -2 dB SNR and 1062 in quiet). Thus N on average is about $2121/56 = 37.88$. For these 2121 trials, the number of utterances with a single (random) error is 11, as shown in Table 2.1. On average, a listener makes a random error every $2121/11 = 192.63$ trials. Hence, the *rate of random errors* is less than 1/190. Since random errors are uncorrelated across utterances, other CVs should also have a similar error rate.

The corresponding value for the number of trials before a random error is made on average, for /t, k, b, d, g/, is 212, 710, 212, 150 and 303 respectively. The outliers are /k/ and /g/, which have very low random error rate. While this needs detailed further study, it is interesting to note that /k/ and /g/ also have the least error ($\approx 0.8\%$) in the low-noise environment. The obvious explanation is that the random errors we defined are *not* totally uncorrelated across utterances, but modulated by the difficulty of the task. In other words, for some LE utterances, even if they have a 1 in N error, the single error may not be random but may be reflective of a low threshold of the feature which the poorest performing listener could be confused on, when presented in noise. For example, of the 11 errors classified as random for /p/, three (f101pe, m115p@, m118pI) have their single error in quiet and are errorless at -2 dB SNR. The responses were /d/, /n/, /?/ (noise only). Consonant

/p/ is not expected to form a confusion group with these consonants [34, 33] and it is therefore reasonable to assert that the score in quiet will be higher than in noise. Hence, it is safe to say that these are truly random.

On the other hand, the other eight sounds have their single error at -2 dB SNR and are confused with /f,k,k,θ,t,f,t,v/. Since /p-t-k/ is known to be a strong confusion group in noise [34, 33], it seems likely that utterances, with their confusions, have a very high threshold for their perceptual feature, hence they are less robust. With -2 dB of noise, the poorest listener tends to prime for these sounds. In short, these errors may not be totally random; that is, the error rate is correlated with the difficulty of the task. Yet, we can still justify calling these utterances "robust" since they have such a very low error. We would likely gain useful information by studying the errors on these utterances at lower SNRs, i.e. by looking at the LE group at -10 dB, in addition to -2 dB and quiet.

In summary, we conjecture that the true random error rate is actually less than 1/300, as for /k/ and /g/. We do not yet have sophisticated analysis techniques to correctly characterize random errors. It seems likely that percentage error may not be an adequate statistic. A more confident error rate could be stated if the errors are analyzed on the basis of confusion groups, listener differences and difference between the two SNRs that were pooled together for the current study. Over time we expect to discover improved methods of monitoring and controlling for these low-grade random errors.

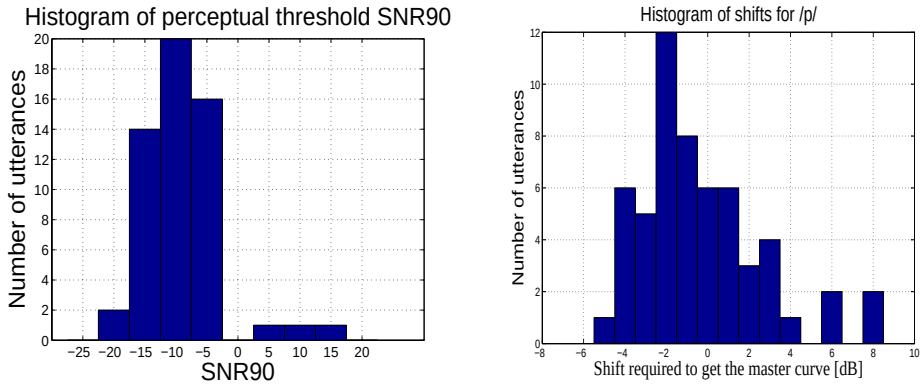
Of course, it is evident that the number of ZE utterances is a function of N, the number of times the utterance is presented in the low-noise environment. The probability of error as a function of N ($P_e(N)$), for large enough N, must become non-zero, due to imprecision in human performance over long periods of trials. For example, even a simple task will have an error for sufficiently

large N. The concept of "zero error" seems essentially flawed, as the number of ZE utterances will tend to zero as N becomes sufficiently large. But, the errors on these utterances would be expected to be of a random nature (low error, high entropy). This is because these sounds are inherently robust (not primable) and have a well-defined perceptual event that is not easily masked. A "zero error" sound implies utterances for which the error (if any) will be of a random nature across thousands of trials, in low additive noise conditions. Its important to note that these sounds make up 63% of the sounds in the MN64 database.

2.6 Modeling the errors: Why does the AI work?

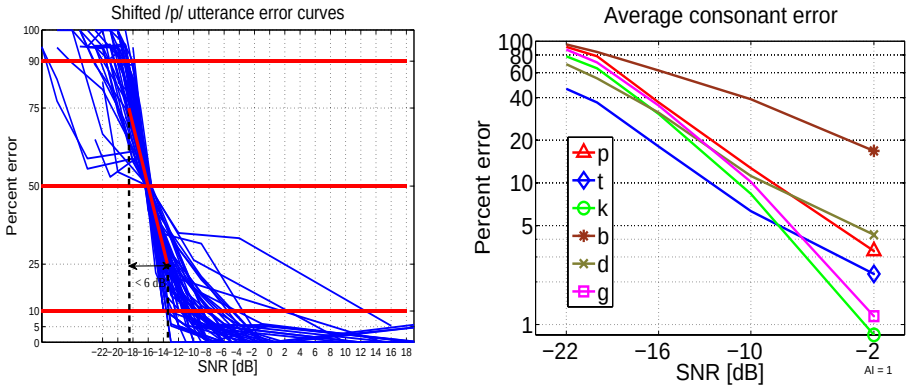
Until now, we have defined -2 dB SNR and quiet as the *low-noise environment* and averaged the data across the two SNRs. Probing further, we see that for the relatively high-salience stop consonants, /t/,/k/,/g/, there is no difference between errors at -2 dB SNR and quiet. The error is about 2% at -2 dB and saturates in quiet. For /p/ and /d/, the error is less than 5% at -2 dB SNR and is reduced by half in quiet. But this is because of a very small number of HE utterances (1 or 2 out of 38) that have errors in quiet. The low salience consonant /b/ has its error at -2 dB SNR (11%) halved in quiet, but the number of utterances having errors in quiet is still significant. Thus, all the stop consonants, with the exception of /b/, have effectively reached their minimum error (e_{min}) at -2 dB SNR. Thereafter the error saturates for 3 out of 6 consonants and is primarily because of 1 or 2 HE utterances for the other two stops. Hence, following the AI model, we map both -2 dB SNR and quiet to AI = 1 (ideal conditions), as documented by French and Steinberg [20].

Of course, it has already been shown by PA07 that the AI model fits these data for C1, C2 and C3. Allen [4] came to a similar conclusion based on the Miller and Nicely [41] dataset. Li and Allen [32] also showed similar results for plosive and fricative groups. In this section, we wish to provide useful insights on the individual utterance error curves, the average of which is an exponential. At very low SNRs, all utterances have 100% error. Hence the average error is also 100%. Similarly, in ideal conditions, most utterances have essentially no error, and the error is e_{min} (due to a small number of erroneous utterances). The individual utterance have different thresholds (Fig. 2.12(a) shows the histogram), but all of them fall from high error ($\approx 75\%$) to low error ($\approx 25\%$) very rapidly, within less than 6 dB. We prove this by aligning all the 56 /p/ utterances at their 50% point to get a *master curve*, the average of these curves. As seen from Fig. 2.12(c), most utterances fall rapidly in the transition from 75% error to 25% error. The slope of the average (master) error curve is about 9.1% per dB. Outliers need further analysis and will be addressed in a future study. The shift required for the alignment is the parameter that is representative of the perceptual threshold of the utterance (ΔSNR), as shown in Fig. 2.12(b). Hence, at a given SNR, the utterances are either at 100% error or at 0% error, with very few utterances in the transition region (since the transition region is so narrow). In other words, an individual utterance error curve approximates a *step function*. The implication of this is that NH speech perception, for robust utterances, is a *binary decision making process*, in which errors are essentially zero above their threshold. The exponential nature of the average curve is due to the threshold distribution. The curve saturates at the ends of the AI range. This saturation as shown in the master curve is similar to Fig. 21 from French and Steinberg [20].



(a) Histogram of the perceptual thresholds of the 56 utterances of /p/ with a bin size of 5 dB.

(b) Histogram of the shifts required to get the "master error curve" for /p/ with a bin size of 1 dB



(c) Individual /p/ utterance errors shifted to align the 50% point at -16 dB.

(d) Average log error curves for the six plosives.

Figure 2.12: (a) Histogram of the perceptual thresholds SNR_{90} values for /p/ utterances. One utterance f106pI never reaches 100% score and its $SNR_{90} = \infty$. Removing the three outliers with high (> 0) threshold values, the SNR_{90} values have a dynamic range of around 25 dB. This is similar to the 30 dynamic range for average speech as shown by French and Steinberg [20]. (b) Histogram of the shifts required to define the *master curve*, with individual error curves aligned at their 50% error values. (c) Individual /p/ errror curves aligned at their 50% score values. The solid (red) line is the average. Most utterances fall from 75% to 25% error within 6 dB. (d) Average log-linear error curves for the six stop consonants, with AI=1 marked at -2 dB SNR. Log-linear regression fits have correlation coefficients of 0.990, 0.997, 0.981, 0.996, 0.998 and 0.992 for /p/, /t/,/k/,/b/,/d/ and /g/ respectively. The average of these six curves is the $e(SNR)$ of Eq. 2.3.

As shown in Fig. 2.12(d), the average error curves for the six stop consonants are close to log-linear. Note that /p/, /t/ and /d/ form a group with a similar slope, as do /k/ and /g/, with comparable values of e_{min}. Hence, an exponential model (exponentials with the same slope added) fits the average error of these two groups. Exponentials, with the same slope, added together would return an exponential with the same slope.

This simple model explains the AI model's characteristics, as given by Eq. 2.5. The exponential error is a simple consequence of the distribution of errors over a large number of utterances having different thresholds, with all but few utterances having no errors, in the low noise environment. Hence, for stop consonants, only a small number of utterances contribute to e_{min}.

2.7 Conclusion

The key conclusions from this study are as follows:

1. Most stop consonants have essentially zero-error in low-noise environments, the summary of which is provided in Fig. 2.11. The consonant /b/ has the smallest ZE group (11/56).

2. NH speech perception for salient syllables is a binary decision making process (you either hear the cue or not), in which the errors are essentially zero when the syllable event is above threshold. This was first shown by [57] for /t/ and is established here for other plosives bases on this complete error analysis on an utterance basis.

3. High error (HE) utterances, due to talker mispronunciation, can be separated from the low error (LE) and medium error (ME) utterances, based on the percentage error and normalized entropy.

4. The source of errors in ambiguous HE stop consonants can almost always be easily explained, using the AI-gram, in terms of the robustness of their perceptual feature, and the feature of the main confusion (conflicting cue) as seen from Figs. 2.3 and 2.6.

5. The average error is exponential with SNR, as modeled by AI theory. This theory works because of the underlying distribution of utterance thresholds, which renders the average error exponential as in Eq. 2.5.

6. The minimum error (e_{min}) under ideal conditions ($AI = 1$), is explained by errors in a small number of highly confusable tokens. These sounds may be characterized by their high SNR_{90} thresholds, typically > 0 dB SNR, or even ∞, for utterances that never reach 90% score.

2.7.1 Implications to ASR

The key issue with automatic speech recognition (ASR) is its fragility due to noise. A confusion matrix (CM) analysis by Sroka and Braida [62] showed that ASR systems (based on different front ends) did a reasonable job in recognizing syllables degraded by lowpass and highpass filtering; however, for syllables degraded by additive speech-shaped noise, none of the automated systems recognized consonants like humans. The phone classification accuracy in ASR systems is only about 82% in quiet [23]. For humans, the score in quiet is commonly assumed to be near 98-98.5% [4]. But again, this is an average over a large number of utterances. Given our present results, we have raised the bar to match human performance. For human speech recognition (HSR), the error is essentially zero. It is the events which make human speech recognition highly robust to noise, as compared to machine recognition [36]. Given precise knowledge of human speech decoding, it must be possible to

exploit this knowledge and build robust ASR front ends that are human-like in performance. Scharenborg [59] gives a comprehensive argument in favor of using the knowledge from HSR research to improve ASR systems.

2.8 Limitations and future work

We believe that this is the first approach in analyzing NH perception of individual utterances, gaining insight into the distribution of errors toward understanding why AI theory works. The work presented is a first step towards this goal and is limited by the simplicity of the analysis. In the future, we wish to carry out an extensive study on understanding the full nature of errors of several other isolated syllables (fricatives and nasals), and also study vowels. We will need a more extensive analysis to fully characterize the utterances in the ME group. Confusion studies and normalized entropy may be the proper tools for such an analysis. In the future, we wish to build a better model of the AI which includes the random error and listener biases. We must also characterize the underlying distribution of each consonant's set of perceptual thresholds.

CHAPTER 3

HEARING-IMPAIRED SPEECH PERCEPTION

To understand HI speech perception, we have been running several psychoacoustic experiments on HI listeners over the past year and have obtained valuable data from about 46 HI ears with mild-to-moderate cochlear hearing loss. The following are some of the tests carried out on a HI listener in multiple sessions:

1. Initial hearing screening: We measure pure tone thresholds of each impaired ear, tymponametry and air-bone conduction gap. A subject is recruited only if he/she has mild-to-moderate cochlear hearing loss, which implies pure tone average (PTA) less than 40 dB, type A tymponametry and no air-bone conduction gap.

2. Middle ear measurements: The reflectance of the middle ear is measured using Mimosa Acoustic's MEPA (Middle ear power analyzer) system. Typically, a person with SNHL is expected to have normal middle ear status.

3. CV discrimination test without amplification: 16 nonsense syllables are presented in speech-weighted noise in this closed-set recognition test. The presentation level is set to the *most comfortable level* (MCL) for each ear.

4. CV discrimination test with amplification: The same CV test is repeated but with amplified sounds, again at MCL. The amplification

scheme used is NAL-R, which prescribes gain equal to half the hearing loss at a particular frequency.

5. Psychophysical tuning curves (PTCs): These are measured at 1, 2 and 4 kHz to assess the tuning and frequency selectivity of the ear at these frequencies. This test is capable of detecting isolated cochlear dead regions.

6. Comodulation masking release (CMR): This experiment (described in detail in chapter 4) is a new methodology to measure cochlear dead regions.

A typical panel for these measurements for a particular HI ear (subject 23 : JT-L) is shown in Figs. 3.1, 3.2 and 3.3. Figure 3.1 shows the PTA, PTC and CMR results.

The other two figures show the confusion pattern of the 16 syllables. A confusion pattern (CP) is a graphical presentation of the scores of the presented syllable (probability of hearing the indicated syllable in the legend given that a particular syllable was presented) as a function of SNR, along with scores of the sounds it is confused with (shown in the legend). The black line (marked with "A" in the legend) is the score for the syllable when it is NAL-R amplified.

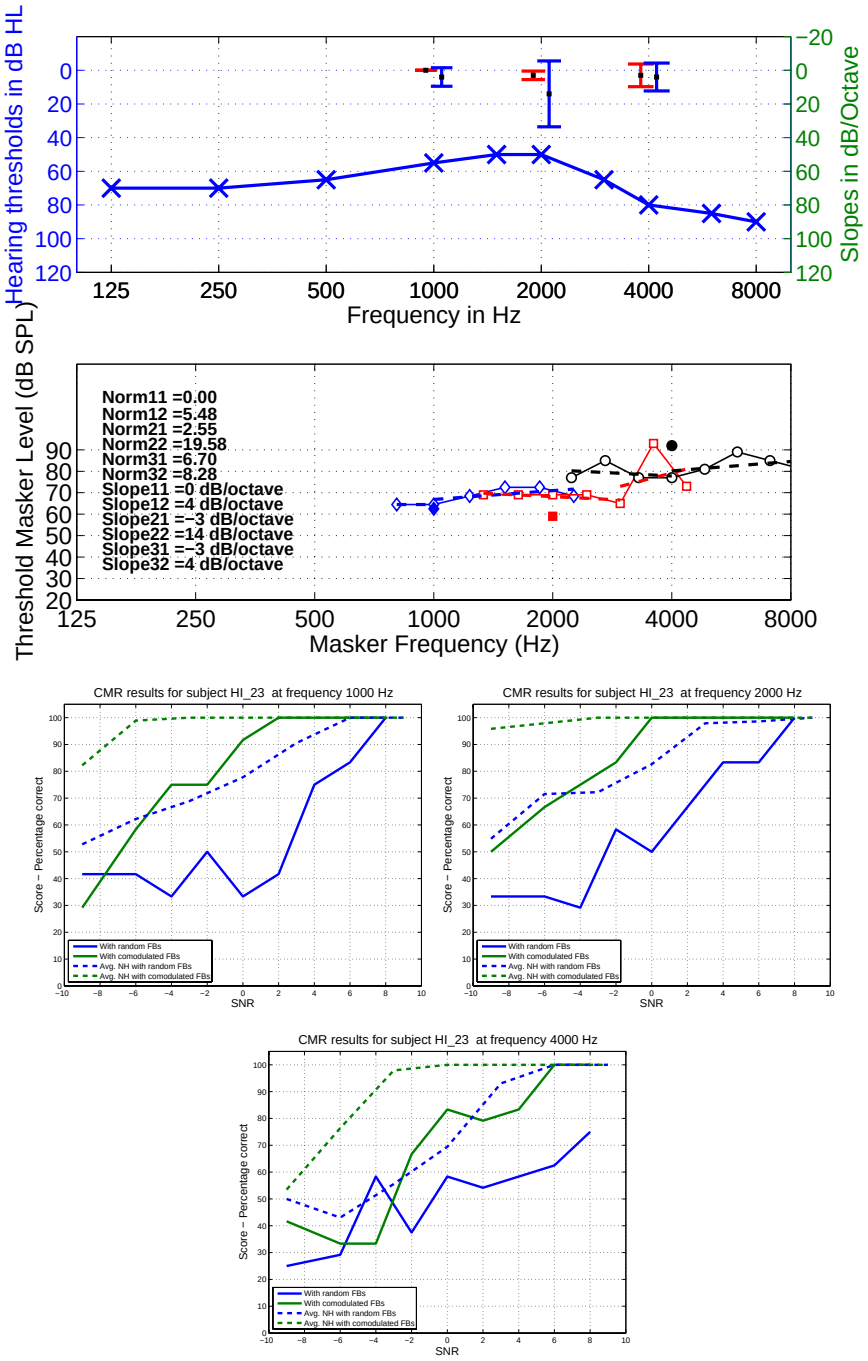


Figure 3.1: PTA, PTC and CMR results resuls for HI subject 23. The top panel shows the hearing thresholds. The panel below it shows the tuning curve at 1, 2 and 4 kHz. These curves are fitted by straight lines whose slope and norm of the residuals (which is a measure of the goodness of fit, smaller value indicates better fit) are indicated on the plot. The residual norm is also plotted on the top panel in red for the negative slope and blue for the positive slope. The bottom three panels show the CMR results, which are explained in detail in Chapter 4.

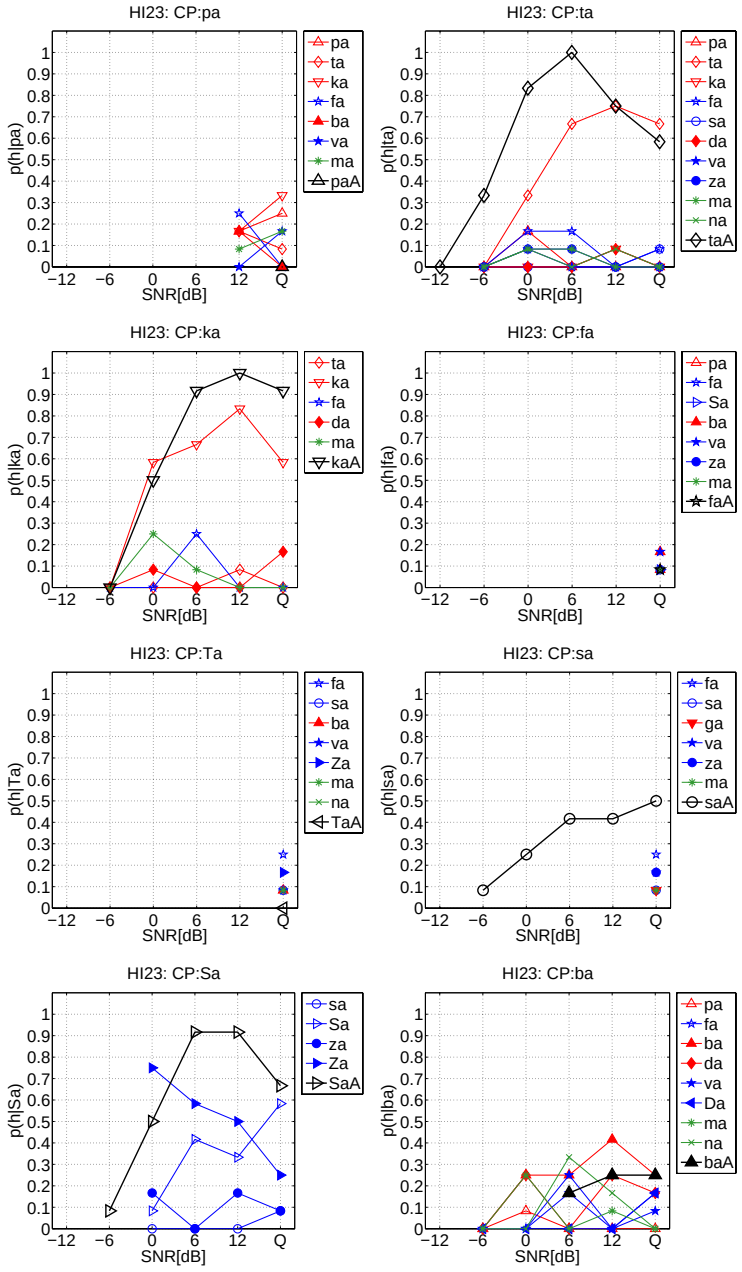


Figure 3.2: Confusion patterns (CP) for HI subject 23 for /pa/, /ta/, /ka/, /fa/, /θa/, /sa/, /ʃa/, /ba/.

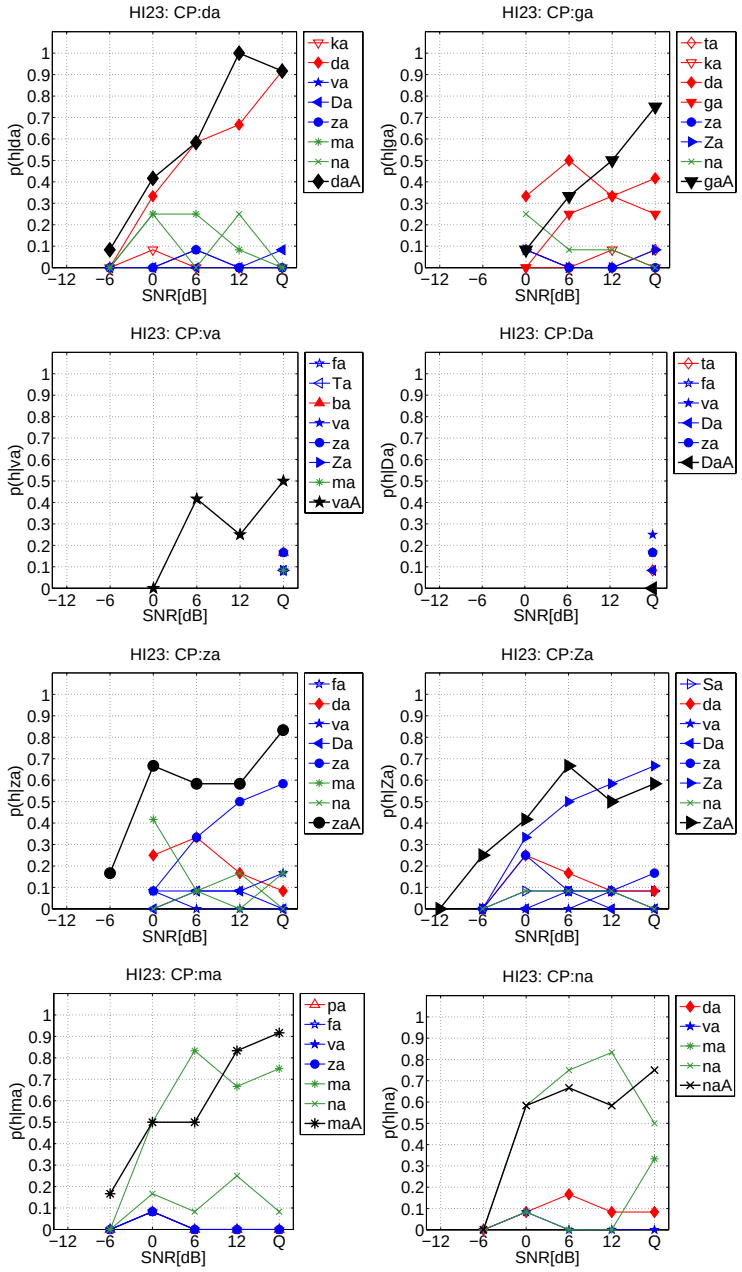


Figure 3.3: Confusion patterns (CP) for HI subject 23 for /da/, /ga/, /va/, /ða/, /za/, /ʒa/, /ma/, /na/.

Our goal is to provide a fundamental understanding of the long unsolved problem of HI speech perception through these several psychoacoustic measurements on the impaired ear.

In the rest of the chapter, we show our results on the CV discrimination test (without NAL-R) and argue that each HI ear is different on consonant perception, and show that there is no clinical measure to capture this dependence.

A patient with SNHL receives recommendations to use a hearing aid or given a profound loss, a cochlear implant, depending on the degree of hearing loss and outcomes from a battery of clinical speech recognition tests. Such assistive listening devices may help the patient hear and even participate in conversations in quiet environments, through appropriate amplification. Unfortunately, most SNHL patients still continue to complain that it remains difficult to understand speech in noisy environments [29, 51, 54, 67].

Hence we argue, based on our present data, that current existing clinical measurements do not reflect the speech perception ability of hearing-impaired listeners, consequently making the listeners less satisfied with the hearing aid, prescribed based on the results of those clinical measurements. We ask the following important questions:

1. How well do the clinical measures predict HI speech perception?

2. How much will current hearing aid fitting procedures benefit every patient concerned?

3. Can existing fitting methods based on outcomes from the typical clinical measures be relied upon to fit modern hearing aids?

We hypothesize that the prescription by either pure-tone audibility or average speech scores, i.e., the *speech recognition threshold* (SRT), results in the SNHL population unsatisfied with their devices. In general, there is a poor correlation between the audiogram, the SRT and individual consonant scores

measured by consonant-confusions, our assumed "gold-standard." Table 3.1 provides the evidence for our hypothesis that all HI ears are different.

3.1 Preliminary analysis: All impaired ears are different

Table 3.1 displays the number of consonant errors in quiet and gives nine examples of HI ears, out of a total of 46 ears. Column ordering is from the easiest (/pa/) to the hardest consonant (/fa/) for our 46 ears. Two syllables /ða/ and /θa/ are not shown in the table because they have high errors (> 40%) even for the average normal hearing (ANH) listeners [50]. Row ordering is from highest to lowest performer of HI listeners, based on average scores. High errors (more than 35% error) are denoted in red in the table. Thus, MA-R is the best performer on average and LB-L is the worst, among the nine ears displayed in the table. Note that the consonant error distribution of the ears is random across the 14 consonants. For example, HI ear MA-R has 50% error on /va/, while there are very few errors in the other syllables. The second best ear TB-R has a spectrum of low-grade problems with /sa/, /za/ and /va/. The third best subject JG-L shows 50% errors on /ba/ and /fa/. Furthermore, subjects with symmetrical hearing loss, MA and LW, have asymmetrical consonant errors. That is, the right ear of MA has 50% /va/ error, but not in the left ear. Instead her left ear has 70% error in /fa/. Subject LW has 42% /ta/ error in only her left ear. These subjects subject demonstrates that pure-tone audibility fails to predict consonant errors. On the other hand, subject LB, with suspectd drug-induced impairment, has a similar error pattern between left and right ear, except for the /va/ syllable. Thus, each HI ear has a different profile, not well represented by any averaged

Table 3.1: Each entry represents the number of errors out of the total number presented. Consonant errors from 9 of 46 hearing-impaired ears in quiet are displayed as an example. The columns are ordered by consonant difficulty from the easiest consonant (pa) to the hardest (fa). Rows (4-12) are ordered by HI performance from the best performer to the worst. First row indicates presented syllables, second row is number of good tokens (tokens on which ANH error is zero) of each syllable, and the third row is the error made by ANH on the tokens (essentially zero). Entries in red indicate significantly high errors ($\geq 50\%$).

	/pa/	/da/	/ma/	/sa/	/ta/	/na/	/ka/	/ba/	/sa/	/za/	/ga/	/va/	/za/	/fa/
# tokens	12	12	12	12	12	12	12	8	8	8	12	12	10	10
ANH	0	0	0	0	0	0	0	0	0	0	0	0	0	0
MA-R	0	0	0	0	0	0	0	0	0	0	0	6	1	2
TB-R	0	2	0	0	0	0	0	0	2	2	0	2	1	1
JG-L	0	0	0	0	1	0	0	4	0	0	2	2	0	5
TB-L	1	2	3	0	0	2	0	0	0	0	0	2	0	4
MA-L	0	0	0	0	1	0	0	0	0	2	1	3	1	7
LW-R	0	2	0	1	1	2	3	0	0	0	3	0	0	4
LW-L	0	2	0	0	5	0	2	0	0	2	0	0	2	7
LB-R	1	0	0	2	0	1	12	1	0	0	5	2	1	10
LB-L	1	0	1	1	0	0	11	1	0	1	5	6	0	9

score. Normal hearing ears have no errors for these sounds, even up to 1000 trials.

The tabulation of the raw data demonstrates how differently the HI ears perceive consonants, which cannot be predicted by (are uncorrelated to) the pure-tone audibility and average speech score. In the following sections, we will review two categories of clinical audiology tests and discuss their advantages and limitations. We raise strong doubts about the reliability and utility of these existing clinical proceedings in predicting *speech loss* in HI listeners.

Our findings are in agreement with those by [51], which includes test-retest data for 6 hearing impaired ears in a gap of about 1 year. They reported good test-retest reliability. The study included 26 HI ears in total. One major goal of our experiment was to raise the number of ears tested (from 23 to an additional 46 ears).

3.2 Reliability of two clinical tests

In this section, we compare our confusion matrix (CM) data to two standard clinical tests. First is the *pure-tone audiometry*, which is ubiquitously used in clinics to measure hearing sensitivity and to determine degree, type and configuration of every hearing loss. Since this measurement is fast, reasonably accurate, and easy to use, it is widely adopted in otology and audiology clinics [61]. It characterizes the threshold of hearing sensitivity (dB hearing level, or dB HL) as a function of frequency from 125 Hz to 8000 Hz in one-octave and optionally 1/2 octave steps, and can establish whether a patient has conductive hearing loss (i.e., middle-ear damage) or SNHL (i.e., cochlear

or auditory nerve damage) by measuring both air and bone conduction. It takes less than 15 minutes per ear. Regardless of these positives, the pure-tone audiometry has a low correlation to a HI listener's perception ability [21], especially for non-flat hearing loss.

In the past, many studies have focused on developing accurate formulae to predict the listener's speech intelligibility from the pure-tone sensitivity. However, a three-frequency average of hearing thresholds at the most important frequencies (i.e., 0.5, 1, and 2 kHz [17]; called *pure-tone average* or PTA) presents a huge individual variance that depends on hearing sensitivity and audiometric configuration [61]. In terms of the hearing sensitivity, normal-to-mild SNHL listeners show a high correlation between audibility and speech perception, whereas moderate-to-severe SNHL listeners have a poor correlation [16]. In terms of audiometric configuration, pure-tone audiogram works best for predicting speech perception of HI listeners with a flat audiogram [11, 17], but works poorly for listeners with a high frequency ski-slope hearing loss [11]. In addition, even though the pure-tone audiogram may be useful for predicting the HI speech perception in quiet, it does not work well when environmental noise increases [17, 61]. The SNHL individual who has *outer hair cell* (OHC) damage typically has elevated audiometric hearing thresholds and low average speech perception scores. Unlike OHC loss, dysfunction of *inner hair cells* (IHCs), denoted as a *cochlear dead region* (CDR), may not be reflected in either loss of audibility or SRT [45, 65, 31]. Thus, based on the phenomenon of *off-frequency listening* [48], Moore and colleagues suggest the use of *psychophysical tuning curves* (PTCs) in order to prevent the pure-tone audiogram measured in quiet from being misread [44]. While the PTCs measure may work, it is not appropriate in the clinic, due to its complexity. As a convenient alternative to the PTC, the *threshold-*

equalizing noise (TEN) test, has been developed [46]. However, recent studies cast doubt on the reliability of the TEN test [38].

Second is the *speech recognition threshold* (SRT). Plomp defines SRT as the speech thresholds at the 50% recognition score [55], when syllables, words, or sentences are presented. Although it is a common and popular clinical measure, the SRT has *four* main limitations. First, clinicians typically use 20 homogeneous high-context *spondee* words (two stressed syllables, e.g., airplane, birth-day, cow-boy) in order to determine the patient's SRT [7], since the spondee words are easier and faster to administer than sentence speech materials [11]. In practice, the clinician usually presents spondee words without the presence of noise [7]. Due to context, such results obtained from only quiet condition are limited in predicting a patient's "real" speech intelligibility. Thus, a clinical *quiet SRT* measure might provide only partial information about HI listeners' speech perception. Second, the SRT considers "average speech scores," instead of an individual consonant score having important and detailed information about perceptual features of speech stimuli that are being misread, due to poor time and frequency resolution/cochlear dead regions/elevated thresholds etc. As a result, the average score, marked in dB, does not characterize the wide variance of inhomogeneous speech intelligibility among HI listeners as shown in Table 3.1 [11]. A third failure of the SRT is its insensitivity to talker dependence, due to the use of a single trained speaker. In the clinic, patients are typically tested with words spoken live through a VU meter of an audiometer by the testing clinician, using the adaptive procedure having 5-dB step size [7]. If the clinician clearly presents the words, the patient will score better. However, if the clinician has an accent or misarticulates, the patient will find it hard to understand the words. Therefore, the patient's SRT result changes depending on the clarity of the

presented words. Lastly, the SRT can be affected by the patient's word bias (e.g., familiarity). American Speech-Language-Hearing Association (ASHA) guidelines recommend that the clinician verify the familiarity of the word lists before starting the SRT measure. If the patient is unfamiliar with some words, they must be removed from the list. The clinician, then, tests the SRT using all familiar spondee words [7]. Such a procedure allows the patient to have a short-term memory of a testing word, and the patient may guess or prime, even when they are unable to hear it. In short, the SRT measures what a patient understands rather than what he/she can actually hear.

The purpose of the present study is to find the most robust measure of HI speech perception. A detailed and scientific measure of consonant-vowel (CV) syllables is needed to reflect the accurate speech perception of HI listeners and offer better insight into the disability that HI listeners experience in everyday listening situations. We expect that our "gold-standard" CV syllable measure, will predict the speech intelligibility of the individual HI ear and will be supported by three main hypotheses: (H1) Compared to ANH listeners, CV syllable scores of HI listeners are significantly poor. (H2) Each HI ears has its own consonant-loss dependence (i.e., heterogeneous consonant score), suggesting that average speech scores are meaningless. (H3) Left and right ears in listeners with a symmetrical hearing loss have unsymmetrical consonant scores, suggesting a serious limitation of the pure-tone audiometry.

Table 3.2 lists the 46 HI ears with their consonant recognition threshold (CRT) and pure-tone average (PTA, average hearing threshold of .5, 1, and 2 kHz). The CRT is the SNR required for 50% corrrect score on average (over 16 consonants). The table is sorted according to the CRT values, from smallest (best ear on average) to the largest (worst ear on average).

Table 3.2: The table contains the CRT and PTA values for the 46 HI ears. The rank ordering is from the lowest CRT (best ear on average) to the highest CRT (worst ear on average). Infinity (∞) value for CRT indicates that the listener had greater than 50% error even in the quiet condition. The CRT values are in dB SNR and the PTA values are in dB HL.

Ear	CRT	PTA	Ear	CRT	PTA
MA-L	-5	41.7	DN-R	1.5	26.7
MJ-R	-4.5	41.7	BD-R	2.5	35.0
MJ-L	-4.5	41.7	LB-R	3	10.0
MA-R	-4	40.0	JS-R	3.5	51.7
SN-R	-3.5	1.7	DG-L	3.5	30.0
JG-R	-3	28.3	EG-L	4	8.3
EM-L	-3	15.0	CP-L	4.5	26.7
TB-L	-3	41.7	DG-R	4.5	55.0
SN-L	-2.5	3.3	LB-L	4.5	11.7
JG-L	-2.5	26.7	WM-R	5	16.7
TB-R	-2.5	46.7	VS-R	5.5	28.3
HV-L	-2.5	41.7	CP-R	6	53.3
BG-L	-2	10.0	VS-L	7	31.7
BG-R	-2	15.0	CL-L	9	63.3
HV-R	-2	38.3	MC-L	9.5	58.3
LW-R	-1	33.3	KW-R	10	15.0
LW-L	-1	40.0	JS-L	10.5	63.3
MB-R	0	18.3	AW-R	12	73.3
PB-L	0	21.7	AS-R	14.5	46.7
PB-R	0.5	23.3	CL-R	16.5	60
MB-L	1	21.7	ES-L	17.5	43.3
BD-L	1	38.3	JT-L	∞	56.7
EG-R	1	16.7	ES-R	∞	56.7

3.3 Methodology

3.3.1 Participants

Twenty-seven HI listeners (17 females and 10 males) were recruited from the University of Illinois at Urbana-Champaign and the Urbana-Champaign community. All listeners were native American-English speakers and were paid to participate. They ranged from 21 to 88 years (mean = 54.96 years) in age. Their hearing screening criteria were normal middle-ear status (A-type of tympanogram) and mild-to-moderate SNHL at PTA although they have various etiology causing their hearing loss. Informed consent was obtained from all participants, and the procedure of the study was approved by the Institutional Review Board of the University of Illinois at Urbana-Champaign.

The results of the hearing screening tests varied in terms of the degree and configuration of hearing loss. Of the participants, 21 subjects showed symmetrical bilateral hearing loss, and 4 showed asymmetrical bilateral hearing loss. Two subjects showed unilateral hearing loss. As a consequence, a total of 48 HI ears were selected for the current experiment. Of these, 10 ears were flat with 3 mild, 4 mild-to-moderate, and 3 moderate SNHL. Another 16 ears showed high-frequencies SNHL varying in the degree of impairment, given that 8 were mild, 6 were moderate, and 2 were moderate-to-severe in hearing loss. A mild-to-moderate high frequency hearing loss appears in 11 ears with a ski-slope of either 1 kHz or 2 kHz. Atypical configurations were also included with 2 at a low-frequency hearing loss, 2 with cookie-bite, 3 with reversed cookie-bite (i.e., the opposite configuration of cookie-bite), and 4 with a mild hearing loss with a notch at 4 kHz.

3.3.2 Stimuli

Isolated English consonant-vowel (CV) syllables were recorded by 18 native American-English speakers. The CV syllables consisted of 16 consonants (6 stops (/p, b, t, d, k, g/, 8 fricatives /f, v, s, ʃ, z, ʒ, ð, θ/, and 2 nasals /m, n/; [41]) followed by the /a/ vowel. The stimuli were selected from the Linguistic Data Consortium (LDC) 2205S22 database [19] and were digitally recorded at a sampling rate of 16 kHz. They were presented in quiet and at five different SNRs (+12, +6, 0, -6, -12 dB) in *speech-weighted noise*. The presentation level of the syllables was adjusted to be at the *most comfortable level* (MCL) for each subject and was SNR dependent, using an external TDT PA5 attenuator. The specific attenuator setting was maintained for each listener throughout the experiment.

3.3.3 Procedure

The test procedures for the CV measurement were similar to those used in the study by Phatak et al. (2009). All subjects had one practice session with 10 syllables in quiet before the experiment. They were asked to identify the consonant in the presented CV syllable by selecting one of 16 software buttons on a computer screen, each labeled with an individual consonant sound. A pronunciation key for each consonant was provided below its button, thus avoiding a possible confusion from any orthographic similarity between consonants. The subjects were allowed to hear the syllable a maximum of three times before making their decisions. After they clicked their response, the next syllable was automatically presented following a short pause.

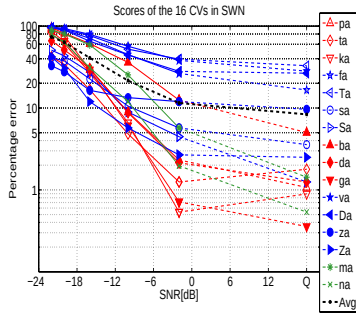
The syllable presentation was randomized over consonants and talkers at each SNR. The subjects were tested in one session, allowing them to have

several breaks. The participants were always tested from easiest to hardest conditions: quiet condition first and then following with a +12 dB to -12 dB SNR condition. A total of 1152 tokens were provided (16 consonants $\times$ 12 presentations $\times$ 6 different noise conditions). However, adaptive procedure was applied with respect to the scores. If the correct score for certain consonant was equal to or less than 3/16 (or 18.75% or three times chance performance), that consonant was not presented at subsequent lower SNRs. The 'noise only' button was allowed if the participant heard only noise without hearing any speech sound. The experiment took 1.5 to 2 hours per the ear and its results were automatically saved in the computer.

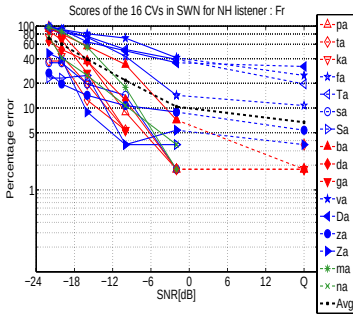
3.4 Results

3.4.1 Comparison of speech scores of ANH to HI listeners

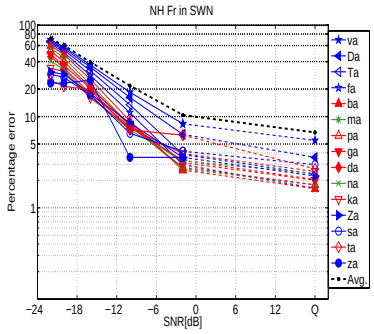
As a control group, data of 10 normal hearing listeners from a previous study [50] was averaged and compared to current data of HI listeners in the same speech-weighted noise. ANH results with 16 CV syllables are summarized in Fig. 3.4, organized into four sub-figures. All sub-figures indicate percent error (log-scale) as a function of SNR. Percent error (P_e or 100% $-$ percent correct) is the ratio of the number of tokens perceived incorrectly to the total number of tokens presented at a given SNR, in percent. Colored lines indicate the percent error of each consonant in the sub-figure, while the average consonant error is marked by a black-dashed line. As noise increases (or lower SNR), all sub-figures of Fig. 3.4 show higher consonant error. Depending on consonant characteristics, however, some of them show much higher error and saturate to chance performance much faster than others. For example, compared to



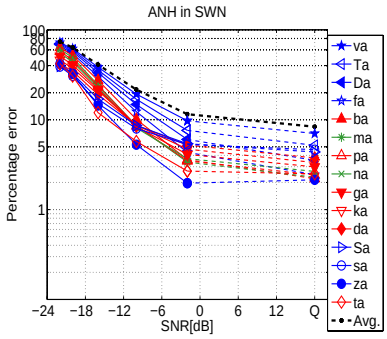
(a) ANH Percent Error of 16 consonants.



(b) NH Percent Error of 16 consonants.



(c) NH Residual Error of 16 consonants.



(d) ANH Residual Error of 16 consonants.

Figure 3.4: Consonant error of one normal hearing listener and average of 10 normal hearing listeners in speech-weighted noise. The legend on the right gives the order of the consonants, from the largest error to the smallest. The rate at which the curves decrease indicates how the total residual error drops as the next high-error sound is removed from the set. This technique is necessary to reduce the otherwise large statistical variability due to the necessarily small sample size. Due to the lack of IPA symbols in Matlab figures, /ʃa/, /ʒa/, /ða/ and /θa/ have been denoted by Sa, Za, Da and Ta respectively in the figure.

/pa/, /ta/ and /ka/, syllables /ða/, /za/ and /ʒa/ show higher errors as a function of SNR.

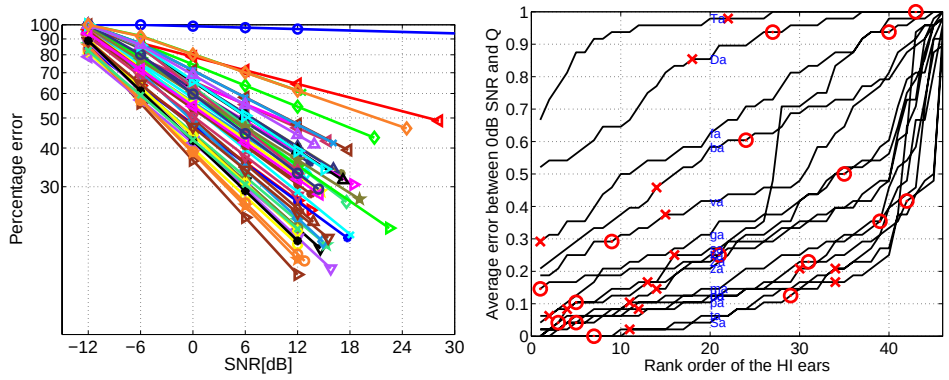
In the left top panel Fig. 3.4(a), the average consonant recognition error (black-dashed line) of ANH listeners increases from 10 to 70 % in quiet to -12 dB SNR. However, this average fails to explain the huge variance among the 16 consonants. From the SRT point of view, *consonant SRT* (SNR at the 50% consonant recognition score [51]) is -18 dB, whereas for /θa/ it is -6 dB and for /ʃa/ it is -22 dB. In the right top panel Fig. 3.4(b), one particular normal hearing (NH) listener is displayed. Regardless of having normal hearing, the listener had a large consonant-error variance, indicating NH consonant-dependence which is statistically similar among the 10 NH listeners. Since the upper two sub-figures look messy (because of small N) and might give difficult and complicated interpretation, new graphical analysis is applied for the study. Percent errors are re-analyzed in terms of consonant error order.

In Fig. 3.4(c) and (d), the legend on the right gives the order of the consonants, from the largest error to the smallest. The rate at which the curves drop indicates how the total residual error drops as the next high error sound is removed from the set: average of 16 consonants for the black-dashed line, average of 15 consonants for right below the black-dashed line with the highest error CV removed, average of 14 for next below with the top two highest error CVs removed, 13 for next, and so on. This graphical technique is necessary to reduce the otherwise large statistical variability due to the necessarily small sample size in the case of individual HI eras. We cannot average scores across HI ears due to large differences between the scores. This technique is applied for the HI results and the data for each HI ear is superimposed onto the ANH data in gray lines as discussed in the following section. We call this procedure "confusion pattern knock-out plot."

3.4.2 Consonant dependence of HI ears: 5 sub-categories

The average error as a function of SNR, for all impaired ears, is shown in Fig. 3.5. Two of the 48 ears failed to complete the experiment and have been discarded for the current analysis. The average of the 16 consonant percent error are parameterized by fitting straight lines as shown in Fig. 3.5(a). The slope of the average log-error curve had a range of 30.72 (log % per dB-SNR) for the best ear to 0.77 for the worst ear. Although some impaired ears look clustered, the distribution is almost uniform over the 46 ears, with the exception of the worst ear, which can identify almost nothing. Unlike the average consonant score per ear, each consonant is separately plotted based on the rank-ordered impaired ear: the best ear is in the left side of x-axis and the worst ear is towards the right (Fig. 3.5(b)). One particular HI ear, TB-L, is marked by x and another ear, PB-R, is shown by o. For instance, TB-L is the highest performer in /fa/ among the HI group, whereas it is located in the middle for /θa/ and /ðа/, and low for /pa/ and /ʃa/. This proves that TB-L has her own consonant-dependence although she is ranked as a good performer in the average consonant score.

As the next step, HI subjects are classified into five groups using *K-mean Cluster Analysis* (K=5) based on the degree of slope of the percent errors versus SNR. Each group shows a significant difference over the SNRs [$F(4,41) = 36.591$, $p < 0.00$ for -12 dB; $F(4,41) = 19.868$, $p < 0.00$ for -6 dB; $F(4,41) = 165.384$, $p < 0.00$ for 0 dB; $F(4,41) = 162.009$, $p < 0.00$ for +6 dB; $F(4,41) = 167.786$, $p < 0.00$ for +12 dB; $F(4,41) = 68.366$, $p < 0.00$ for quiet condition]. Groups 1, 2, 3, 4, and 5 were defined as the best (15 ears), high (13 ears), medium (13 ears), low (4 ears), and worst (1 ear) performance groups, respectively. More detailed explanations of each group follow.



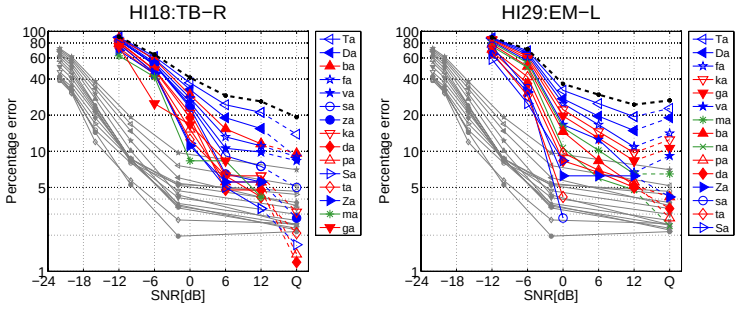
(a) Parameterized analysis of average percent error.

(b) Consonant error of rank-ordered HI ears.

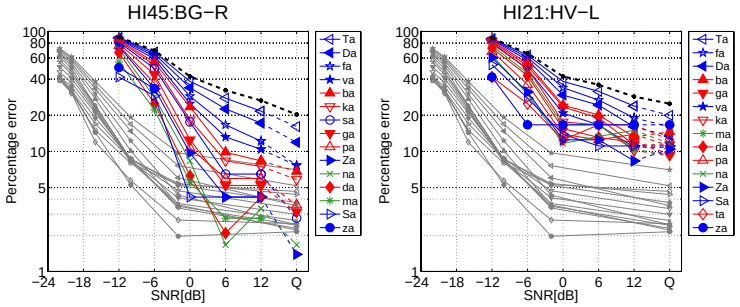
Figure 3.5: (a) Parameterized analysis of average percent error (%) for 46 HI listeners based on the slope of their average log-error curve. (b) Consonant error of rank-ordered 46 HI ears: abscissa indicates HI listener order, from best to worst for each CV and the ordinate the is average consonant error between 0 and quiet in 16 CV. Two particular impaired ears are marked by symbols O and X.

Group 1: 15 Best Performance Ears in Average Scores • 15 HI ears are clustered into the best performance group. Among them, four impaired ears are selected. The data for each HI is superimposed onto the gray-lined ANH data (Fig. 3.6). These four ears have a similar pattern of the average error in the top black-dashed line: about 20% error in quiet, 40% error at 0 dB, and saturated to 100% at -12 dB. In addition, all four ears show highest errors in /θa/ and /ða/. However, they differ in overall order of consonant difficulty and the variance of consonant errors. HI ear TB-R shows the largest variance (i.e., 1∼20%) of consonants error in the quiet condition, but the variance gradually decreases at -6 dB. Compared to TB-R, HI ears EM-L and BG-R have larger variance at the same -6 dB SNR. On the other hand, although the left ear of HI subject HV is in the same group with the remaining 14 ears and has very similar percent error in average consonant, its variance of consonant errors is much smaller even in quiet condition, showing

high error in most consonants. In Fig. 3.6, TB-R (a) and BG-R (c) have a few consonant errors in normal hearing range. Thus, though the average error is same for all the 15 ears in this group, the ears each have a different mechanism to obtain the same average error. Twelve ears have a larger spread which indicates that they have trouble with only a few consonants. The remaining 3 ears have a tight spread, which implies that most CVs have high errors.



(a) Right Ear of HI subject TB. (b) Left Ear of HI subject EM.

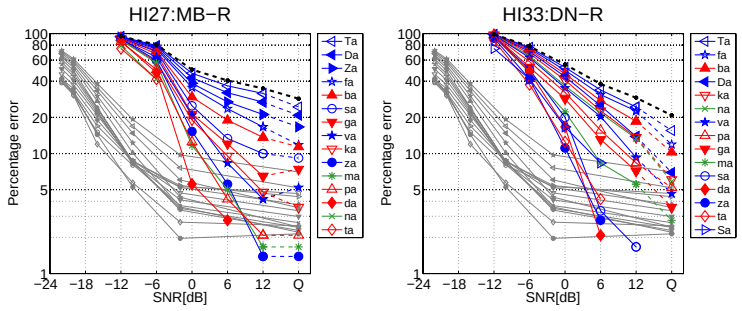


(c) Right Ear of HI subject BG. (d) Left Ear of HI subject HV.

Figure 3.6: An example of four ears selected from Group 1 (best performing ear). Four sub-figures (a)∼ (d) show the consonant errors for four HI ears TB-R, EM-L, BG-R, and HV-L in speech-weighted noise, as a function of the SNR. The data for each HI is superimposed onto the ANH data (gray lines). The dashed top curve is the average error of all 16 syllables. Each dashed curve with a symbol describes the average error of the rest of the consonants. That is, it shows the average loss after removing the worst sound, the two worst sounds, the three worst, etc.

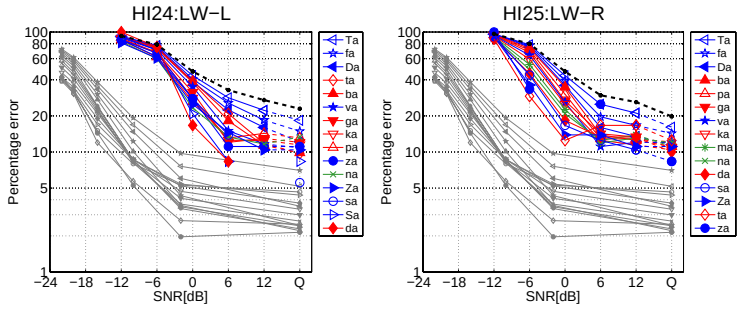
Group 2: 15 High Performance Ears in Average Scores • There are 15 HI ears in the second group, too, the high performance group in

terms of the average consonant scores. Among them, four impaired ears are chosen as the example in Fig. 3.7. Like Group 1, these HI ears are not significantly different each from other in the averaged consonant errors and pattern, excluding the right ear of HI subject DN. However, two upper panels, (a) and (b), show much larger variance than the lower panels (c) and (d) although all four are included in the same performance group based on average scores. In panel (a), /za/ and /ma/ error is smaller than ANH at +12 dB and quiet, and /ta/ is much lower error than ANH even at 0 dB SNR. Overall /ta/, /na/, and /da/ show lower errors than other syllables across SNRs. On the other hand, HI DN-R has 55% average consonant error at 0 dB, and it has a much steeper average error curve (panel(b)). Panels (c) and (d) are the left and right ears, respectively, of one HI subject, LW. Although she has symmetrical hearing loss in pure-tone audiogram, her consonant perception differs in order of difficulty between left and right ear. Again, there is an obvious difference of the consonant-dependence in each impaired ear. Eleven subjects have the pattern having huge variance and the rest of them have less variance.



(a) Right Ear of HI subject MB.

(b) Right Ear of HI subject DN.

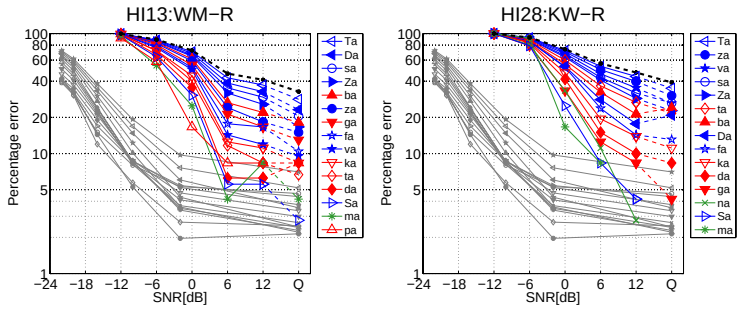


(c) Left Ear of HI subject LW.

(d) Right Ear of HI subject LW.

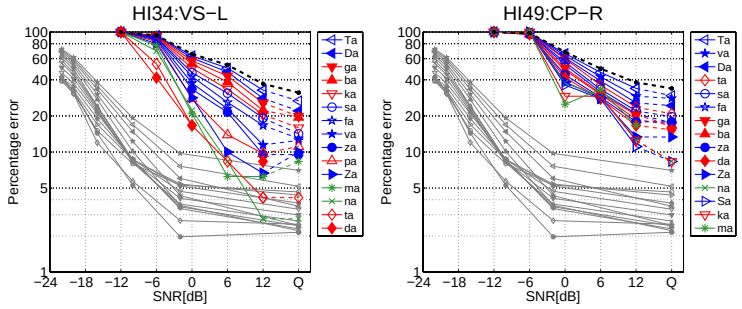
Figure 3.7: An example of four ears selected from Group 2 (high performance ear). The upper two panels show confusion pattern knock-out plots having higher consonant variance than the lower two panels although all four ears are within same group.

Group 3: 13 Medium Performance Ears in Average Scores • In the third (medium) performance group of averaged consonant score, 13 HI ears are included. Among them, four impaired ears are chosen as the example in Fig. 3.8. Compared to the best and high performance ears, the medium performance group is much more sensitive to noise, showing higher error as well as less consonant variance at -6 dB. The right ear of HI subject CP is not different from the other three sub-figures in terms of the average consonant error. However, it shows much less consonant dependence and does not have any consonants with less than 10% error.



(a) Right Ear of HI subject WM.

(b) Right Ear of HI subject KW.

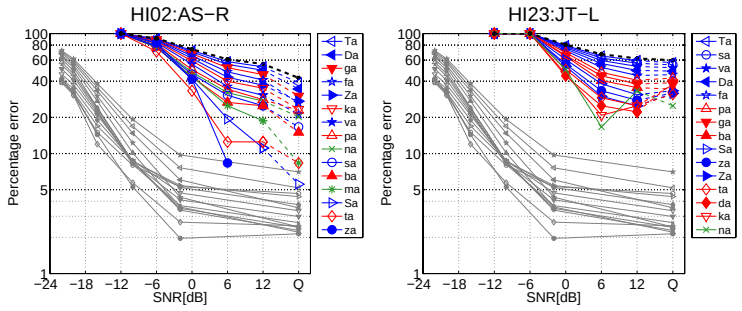


(c) Left Ear of HI subject VS.

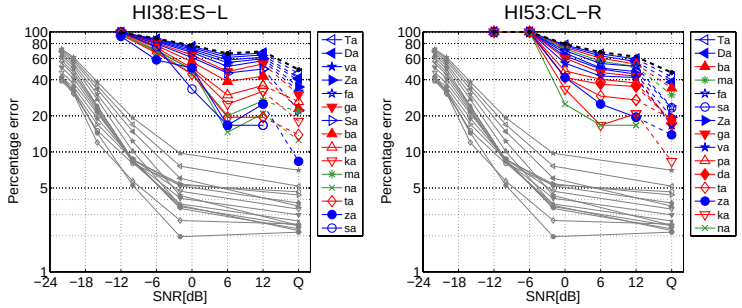
(d) Right Ear of HI subject CP.

Figure 3.8: An example of four ears selected from Group 3 (medium performance ear). The upper two panels show higher consonant variance than the lower two panels although all four ears are within same group.

Group 4: 4 Low Performance Ears in Average Scores • The four impaired ears that belong to the fourth (low performance) group are shown in Fig. 3.9. The average consonant error curve indicated very high error, greater than 40% even in quiet. Although there is a certain range of the consonant dependence at +6 dB, this group displays less consonant dependence compared to other better performance groups. No ear performs better than ANH in this group. HI ears JT-L and CL-R show 100% error for all consonants beyond -6 dB SNR.



(a) Right Ear of HI subject AS. (b) Left Ear of HI subject JT.

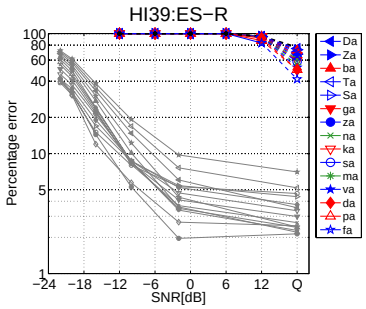


(c) Left Ear of HI subject ES. (d) Right Ear of HI subject CL.

Figure 3.9: The four ears from Group 4 (low performance): The upper two panels show higher consonant variance than the lower two panels although all four ears are within same group.

Group 5: 1 Worst Performance Ear in Average Scores • The worst performing HI ear is displayed in Fig. 3.10. The right ear of HI subject ES shows 40% or higher error and 100% error in all consonants in quiet and +6 dB SNR, respectively. Although he has high frequency hearing loss from 1.5 kHz and 25∼30 dB HL in low frequencies, his consonant perception score is not consistent with pure-tone audiogram due to high error in all consonants.

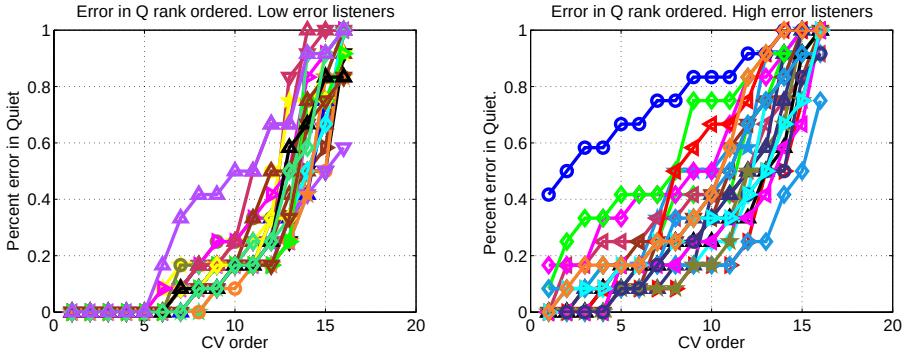
Clearly every HI ear has a different profile, not well represented by any averaged score. Another way of proving the same fact is by rank ordering the listeners on their errors in quiet. As shown in Fig. 3.11, the listeners are rank ordered on the basis of their consonant order. As shown in the left



(a) Right Ear of HI subject ES.

Figure 3.10: An example of a confusion pattern knock-out plot for one ear selected from Group 5 (worst performance).

panel, 24 out of 46 listeners have zero error in at least 5 of the 16 consonants, in quiet. These are the so-called "low error listeners." Conversely, the right panel shows the "high error listeners" (22 out of 46), who have less than 5 out of 16 errorless consonants.



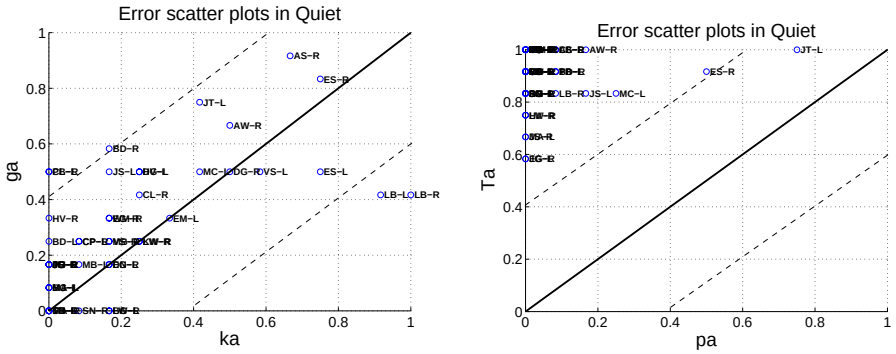
(a) Low error listeners

(b) High error listeners

Figure 3.11: Rank ordering the listeners according to their own consonant dependence.

Also, although not shown in the figure, the rank order is different for each ear, which implies that if a particular ear is a good performer for a particular CV, it may not be the best for other CVs. Of course, scores of consonants having similar frequency region of importance (similar frequency range of the perceptual cues) like /ka/ and /ga/ (mid frequency burst) or ta and da (high frequency burst) are expected to be correlated. For example, Fig. 3.12 shows

the scatter plot of /ka/ vs. /ga/ and /pa/ vs. /θa/.



(a) Scatter plot of /ka/ vs. /ga/

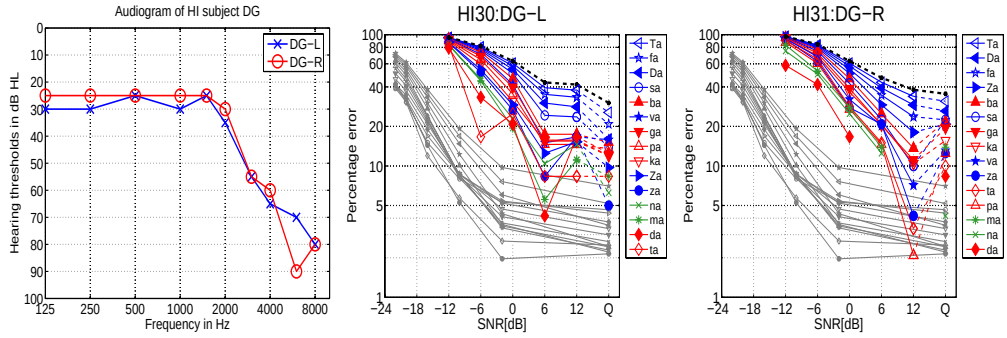
(b) Scatter plot of /pa/ vs. /θa/

Figure 3.12: a) Forty-six HI ears marked with their error in quiet for /ka/ on the x-axis and /ga/ on the y-axis. Both these consonants have a similar perceptual cue in terms of frequency (a mid-frequency burst around 2 kHz) and their scores are expected to be correlated as seen from this figure. b) Forty-six HI ears marked with their error in quiet for /pa/ on the x-axis and /θa/ on the y-axis. Clearly, they are not correlated. In fact, /pa/ is the easiest sound to perceive in the database of the 46 ears and /θa/ is the hardest. Most HI ears have high errors on /θa/, while only two ES-R and JT-L perform poorly on /pa/.

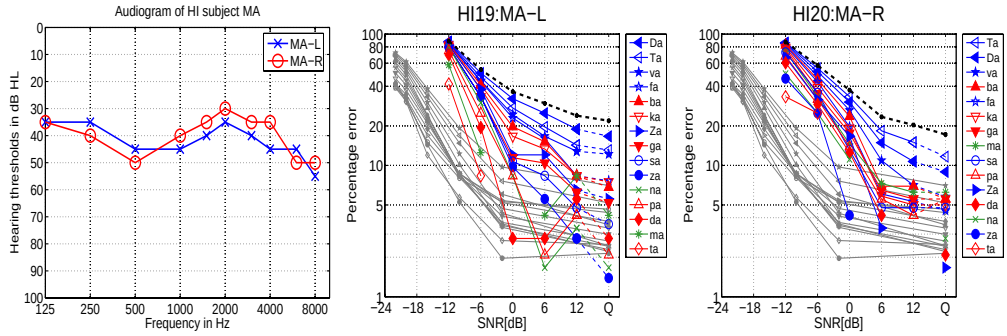
3.4.3 Consonant error difference of symmetrical hearing loss: Left versus right ear

Ten HI subjects having symmetrical hearing loss are analyzed. In Fig. 3.13, the pure-tone audiograms from three subjects are shown. These subjects have similar hearing levels but are different in the consonant scores (middle and right panels). That is, two ears of each subject have functionally identical audiograms, but very different consonant errors. For example, MJ-L and MJ-R (bottom three panels) have a measurable difference with respect to the most difficult consonants. In the left ear, the order of difficult consonants is /ða/, /θa/, /fa/, and /va/. However, it is /θa/, /fa/, /ða/, and /ba/ for the right ear. She has many consonants less than 5% error in the left ear, which might be consistent with her left ear preference for using the telephone,

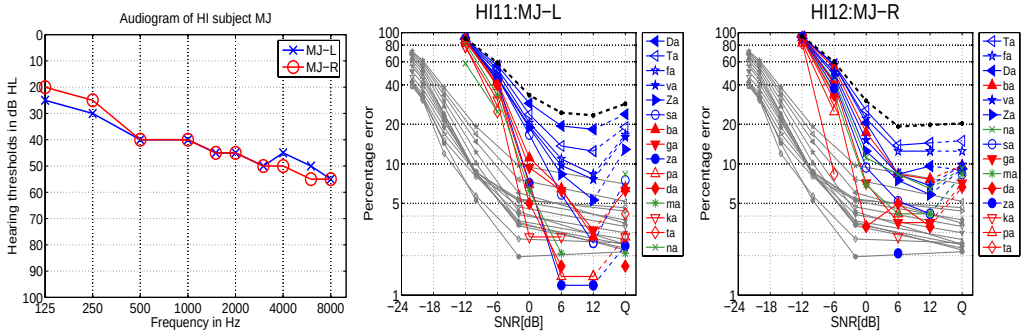
though the average score is better for the right ear.



(a) Audiogram HI-DG. (b) Left Ear of HI subject DG. (c) Right Ear of HI subject DG.



(d) Audiogram HI-MA. (e) Left Ear of HI subject MA. (f) Right Ear of HI subject MA.



(g) Audiogram HI-MJ. (h) Left Ear of HI subject MJ. (i) Right Ear of HI subject MJ.

Figure 3.13: An example of left and right ears difference for three HI listeners having symmetric hearing loss. The left three panels are pure-tone audiogram of three HI subjects, the middle three panels are consonant confusion knock-out percent error of the left ear of the three subjects, and the right panels are from the right ear.

3.5 Discussion

The goal of the present study is to determine the most robust measurement of speech perception for HI listeners among current existing clinical measures. Each of the specific research hypotheses posed earlier is addressed in turn.

Compared to ANH, CV syllable scores of HI listeners are significantly poor. To be exact, most SNHL listeners have both loss of sensitivity and loss of speech clarity [28]. The loss of sensitivity is represented in pure-tone audiogram and can be easily evaluated by degree of hearing loss. However, as Killion pointed out, the loss of speech clarity (or simply called SNR-loss) is not revealed in either the audiogram or SRT measure [28]. Thus, although HI listeners wear hearing aids they still complain about unclear speech in noise because they cannot hear certain perceptual cues due to their hearing loss or the masking effect introduced by the noise.

In our results, most HI listeners show poorer consonant perception in quiet as well as lower SNRs than ANH with respect to the average scores. This result strongly supports SNR-loss of HI listeners and is consistent with Killion's results [28]. As shown in Fig. 3.4, HI ears always have significantly higher error in average consonant scores than ANH along SNRs. However, in the current study, we find that SNR-loss of the HI listener, which is the increased SNR required by the listener to understand overall speech in noise, is recorded in a few, not all consonants. The consonant dependence of ANH listeners does not statistically vary across listeners (Fig. 3.4(c) and (d)), but each HI listener has his/her own profile. HI consonant confusion is much more complicated; some consonants are much worse than ANH, yet others are not. Therefore, we assert that HI ears have SNR-loss as well as consonant dependence and we coin the term *speech loss*, in order to better explain the

combination of the two factors. Speech loss cannot be predicted from any other clinical measure; thus, we suggest that a CV syllable test is needed for HI ears in clinics and research laboratories to be able to diagnose the sounds that the ear misses.

Does a HI listener have his/her own consonant-dependence, suggesting a deficiency in the current SRT measure? Although we do not have data obtained from spondee SRT measurement used in the clinic, our average consonant error (or the consonant SRT) shows a serious limitation for predicting the overall speech recognition of hearing impairment. Again, because HI ears show error in only a few sounds, either average score or word/sentence score is too limited to show their errors. In the Fig. 3.5(b), the consonant error of HI ear TB-L (x mark) shows the diversity. That is, she is a good performer in /pa/, ta/, and /ka/ and the highest performer in /fa/, but bad in /ʃa/ and /da/ syllables. The most difficult consonants for all HI listeners are /θa/ and /ða/, which are difficult even for normal hearing listeners [50]. However, this cannot be predicted from any other clinical measure, as mentioned before. Furthermore, the clinical SRT measured in quiet condition cannot determine a HI listener's speech perception in noise. Thus, we recommend using the CV syllable test, which considers both natural noisy situation and detailed information of consonant perception. We believe that this will provide novel ideas to effectively enhance a HI listener's speech understanding, with better amplification strategies in the future.

Is there a significant difference in consonant scores between left and right ears for listeners with a symmetrical hearing loss? If so, does this imply a limitation of the pure-tone audiometry? According to a previous study, pure-tone audibility has limitations in predicting speech perception because the loss of audibility and loss of speech clarity

(i.e., SNR-loss) are separated in terms of their functions [28]. In other words, there is a major difference between hearing speech (i.e., audibility of speech) and actually understanding it (i.e., intelligibility of speech); some individuals have a much greater loss of ability to understand speech in noise than would be predicted from their audiogram [28]. As a solution for the limitation of pure-tone audiometry, Killion suggests a graphical *Count-the-Dot Audiogram Method* for estimating *Articulation Index* (AI) [47]. Although his method is an easy and practical way to tell us how much clarity the hearing-impaired patient has lost, by computing the number of dots above the audiogram [28], there is still a flaw as it gives no explanation of inhomogeneous HI speech perception. That is, the count-the-dot AI does not explain why HI ears have different scores or large individual variance.

3.5.1 Limitations and future work

We have successfully developed full-rank consonant-confusion matrices as a function of SNR and propose it as the most robust technique to measure speech perception of HI listeners. One limitation of the current method is that it is time-consuming in its current format. By reducing the number of syllable presentations and carefully selecting good tokens, we will work on developing a convenient, fast, as well as statistically proven, speech perception test battery for clinical usage. Results from NH studies (detailed in Chapter 2) show that the majority of utterances in the PA07 database are robust and we must use these sounds to test measure HI speech loss.

In addition, there are a couple of possible future research goals, and we have continued several ongoing studies related to the consonant confusion measures. *First* is to find the relation between consonant error and cochlear

dead region. It may be possible to use the consonant confusion matrices in order to detect the cochlear dead region, instead of spending a lot of time on current existing psychoacoustic measures (e.g., psychophysical turning curve (PTC)), which is not ideal for clinical use. *Second* is to examine the benefit of amplified speech through the individual consonant measures, our gold standard. Amplification (using a hearing aid) over a frequency range corresponding to a dead region may not be beneficial and may even impair speech intelligibility [45]. The study will address the problem of speech perception in the noisy situation. *Third*, we continue to work on establishing a new and delicate amplification formula that is based on individual speech scores, while applying differential amplification, manipulating both frequency and loudness. The long-term goal is to help HI listeners hear their inaudible sounds while not reducing the intelligibility of sounds that they presently hear. This is our "do no evil" strategy. The study could thus be significant in helping HI listeners hear conversations more clearly and further aid in audiological diagnosis and successful rehabilitation in the future.

3.6 Conclusions

To summarize, the key results of the current study are as follows:

1. Although HI ears need higher SNR than ANH in most consonants, they need different SNR depending on consonants; some consonants can, in fact, be in the normal hearing range. The HI ears differ from each other in consonant perception.

2. Regardless of similar loss of audibility and configuration, individuals with hearing loss show different consonant dependence. That is, their average consonant error, or consonant SRT, does not explain the huge variance of

consonant recognition. In this matter, the average scores of consonants (or spondee words) does not characterize each HI listener's consonant profile and thus, does not solve the fundamental problem of their poor speech perception, resulting in low satisfaction with amplification.

3. Individuals who have symmetrical hearing loss (i.e., pure-tone sensitivity) show different consonant-dependence in the left vs. the right ear, denoted *asymmetrical consonant error.* There is a difference in percent error as well as in the ordering of difficult consonants. This information is not reflected in the clinical pure-tone audiogram.

Thus, compared to current existing clinical measures, individual CV syllable recognition is the most robust and accurate measure of speech perception for HI listeners. The syllable measure might give detailed information about characteristics of the HI listeners' speech perception, resulting in increased insight into the problems that HI listeners experience in everyday listening situations.

CHAPTER 4

DETECTING COCHLEAR DEAD REGIONS

Cochlear dead regions are places along the basilar membrane of the cochlea where the inner hair cells (IHCs) are non-functioning due to damaged or missing IHC cilia [43]; which are fine "hair bundles" at the top of the cell that transduce the basilar membrane motions. In addition, the afferent auditory neurons innervating those places may be degenerate [31]. A dead region can be described in terms of the characteristic frequency of the IHC at the specific region on the basilar membrane where it occurs. It is widely accepted that speech perception is seriously degraded in these regions of degraded transduction. Important conclusions from a study on cochlear dead regions by Brian Moore in 2001 [43], are:

1. "Dead regions may be relatively common in people with moderate-to-severe sensorineural hearing loss."

2. "Dead regions cannot be reliably diagnosed from the pure tone audiogram."

3. "Psychophysical Tuning Curves (PTCs) provide a useful way of detecting dead regions and defining their boundaries. However, the determination of PTCs is probably too time-consuming to be used for routine diagnosis of dead regions in clinical practice."

4. "Amplification of frequencies well inside a high-frequency dead region usually does not improve speech intelligibility, and may sometimes impair it."

Another recent diagnostic tool is the TEN (Threshold Equalizing Noise) test which involves measuring the threshold for detecting a sinusoidal tone presented in a special background noise called the *threshold equalizing noise*. However, it has been argued that this test is not accurate [65]. Owing to the limitations of the currently used procedures to detect dead cochlear regions, and the fact that the presence or absence of dead regions can have important implications in fitting of hearing aids, our research group at UIUC has been looking for alternative methods to detect dead regions. One such technique we are evaluating uses the comodulation masking release (CMR) paradigm.

4.1 The CMR effect

The CMR effect was first discovered by Hall, Haggard, and Fernandes in 1984 [22] and is decribed as follows. Consider trying to detect a pure tone (target) in a narrow-band noise (masker). The target detection threshold increases with increase in the noise bandwidth, as long as it is within the bandwidth of the auditory filter centered around the tone frequency, but the masker has no effect for bandwidth wider than a critical band. However, if we extend the masker to *flanking bands*, well outside the critical bandwidth, the target threshold was shown to depend on the correlations between the modulation envelopes of the target and flanking bands maskers. Surprisingly, it becomes easier to detect the tone, in spite of having added more noise, when the target and flanking maskers are correlated! Thus these flanking bands, having coherent modulation, produce masking release (Fig. 4.1 from [53]).

This effect can be attributed to the ability of the auditory system to make comparisons of envelope fluctuations across frequency.

Clearly, it is important to know if this release in masking is relevant to speech perception in comodulated noise. This is an important research question, not properly addressed in the literature. For example, [58] used wide-band gated noise as a masker. Such a noise would qualify as "comodulated," making CMR a relevant experimental paradigm.

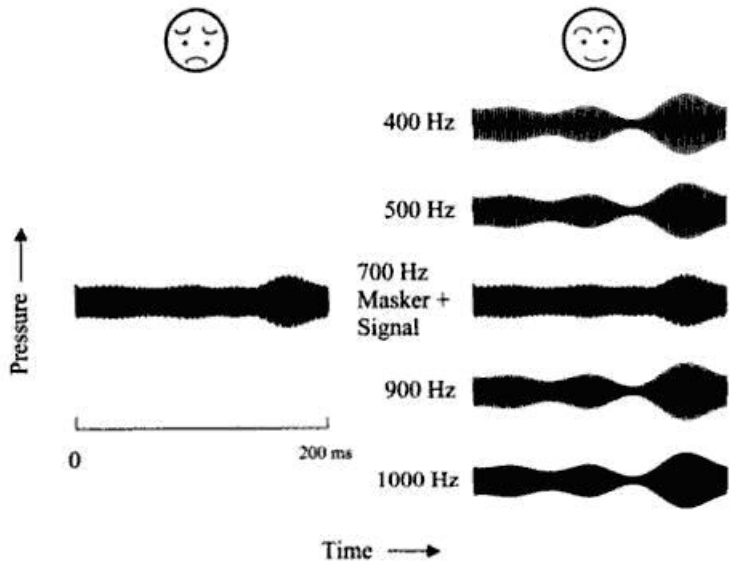


Figure 4.1: A just detectable 700 Hz pure tone is centered on and masked by an envelope-modulated band of noise. When co-modulated narrow-band flanking maskers are added, the detectability increases (from [53] p. 165]).

4.2 CMR as a diagnostic tool

While the CMR effect has been extensively studied by many researchers, the contribution of the present work is in trying to use CMR as a diagnostic tool to detect isolated cochlear dead regions in hearing-impaired individuals. Our

hypothesis is that, due to loss of tuning in a dead region, there would be no release in masking if the target tone is in a dead region. This is because the envelope would not be resolved in a dead region and the consistent pattern of comdulated envelopes in various critical bands, would be broken. Hence, there would be no release in masking. The CMR test can be performed at any frequency by choosing suitable flanking band frequencies.

4.2.1 Methods

The current study includes 19 subjects (33 ears) having mild-to-moderate sensorineural hearing loss. The procedure requires the HI listener to detect a pure tone target in the presence of a narrow masker (on-signal band) with four flanking bands in two conditions: (1) the flanking bands have random amplitude modulations, (2) the flanking bands are comodulated so to have the same envelope fluctuation as the masker. This procedure is conducted at various signal-to-noise conditions [-9 to 9 dB]. The SNR order is randomized and a different noise spectrum is created for each trial. At each SNR and each condition, N (8 or 12) trials are presented to the listener in a three interval forced choice (3IFC) method (chance = 33%). One of these intervals contains the target tone. The listener must identify the interval having the target or can choose "SAME" if all three intervals seem identical. The responses belonging to the bin "SAME" are uniformly distributed into the three interval bins. The listener is allowed to listen to the trial, using a "REPEAT button", a maximum of three times before making his/her decision.

The PTC at 1, 2 and 4 kHz is next measured, using a computer program written in Matlab for this purpose. Finally a full rank confusion matrix (CM) is measured using nonsense CV syllables under various noise conditions, the

procedure of which is described in detail in Chapter 3. Since we know the speech cues for these consonants a priori [34, 40], consonant error patterns from the CM experiment allow estimates of the possibility of a dead region. The hypothesis is that if a subject has a CDR at the frequency location of the critical cue, the ear would miss the critical feature, consequently making that syllable inaudible. We assume that the HI ears use the same perceptual features as the ANH users. This is a reasonable assumption, especially for HI ears who have post-lingual hearing loss. Also, since we use maxEnt syllables, context effects are minimized and a person's vocabulary or linguistic knowledge is irrelevant to the nature of the task.

4.3 CMR results

4.3.1 CMR results on normals (normative data)

Subjects with normal hearing are expected to have a masking release of about 5-7 dB at all measured frequencies [22]. Figures 4.2, 4.3 and 4.4 show the results of individual NH subjects in the CMR experiment at three frequencies: 1, 2 and 4 kHz. The data has been parameterized by fitting it with straight lines (shown as dotted lines in the figures). The green lines are scores with comodulated flanking bands and we see that these scores are always higher than when randomly modulated flanking bands are presented, because of the CMR effect. Figure 4.5 shows the average curves in solid lines with the individual data in dotted lines. Figure 4.6 shows the results of the average scores across NH listeners. The blue line is the real average and the green line is the average of parameterized curves. As seen in the figure, the two curves are similar.

1 kHz : 5 subjects, FY-L, JH-L,RS-R, TK-R, WK-R

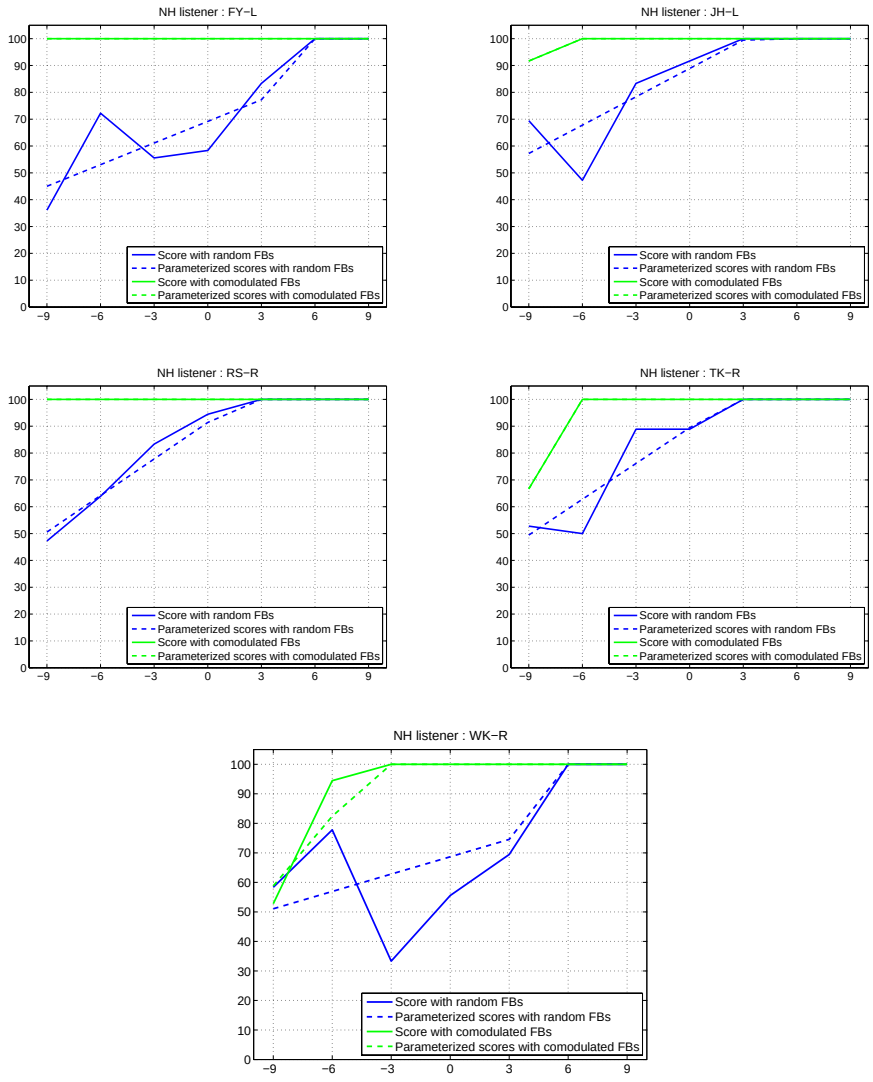


Figure 4.2: CMR results for NH listeners at 1 kHz. The scores for random flanking bands (FBs) condition are shown in blue, while those for the comodulated FBs condition are shown in green. The solid lines are the actual scores as a function of SNR, while the dashed lines are parametric scores obtained by fitting the data with straight lines. For all the five subjects, the score for comodulated case is always greater than with the random case. Hence, all the five ears have CMR.

2 kHz : 4 subjects, FY-L, JH-L, TK-R, WK-R

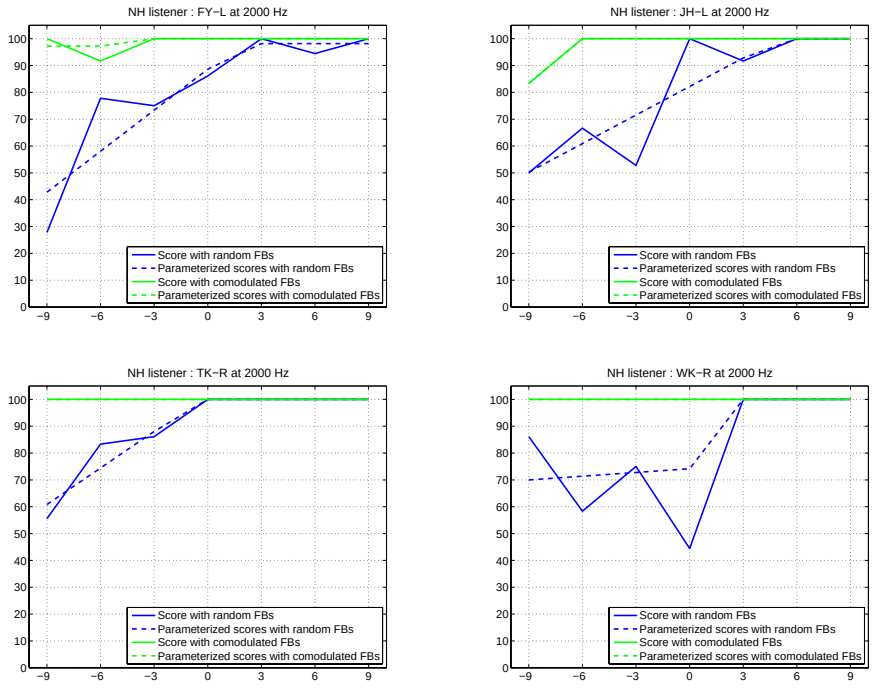


Figure 4.3: CMR results for NH listeners at 2 kHz. The scores for random flanking bands (FBs) condition are shown in blue, while those for the comodulated FBs condition are shown in green. For all the four subjects, the score for comodulated case is always greater than with the random case. Hence, all the five ears have CMR. Hence, they do not have CDRs at 2 kHz.

4 kHz : 4 subjects, JH-L,RS-R, TK-R, WK-R

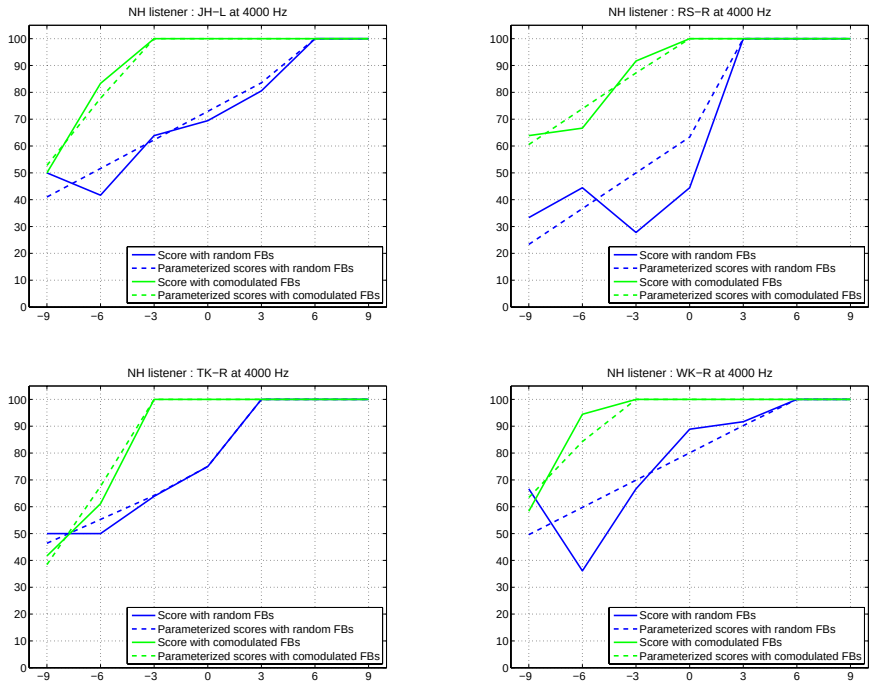


Figure 4.4: CMR results for NH listeners at 4 kHz. For all the four subjects, the score for comodulated case (solid green line) is always greater than with the random case. Hence, all the four ears have CMR. Hence, they do not have CDRs at 4 kHz.

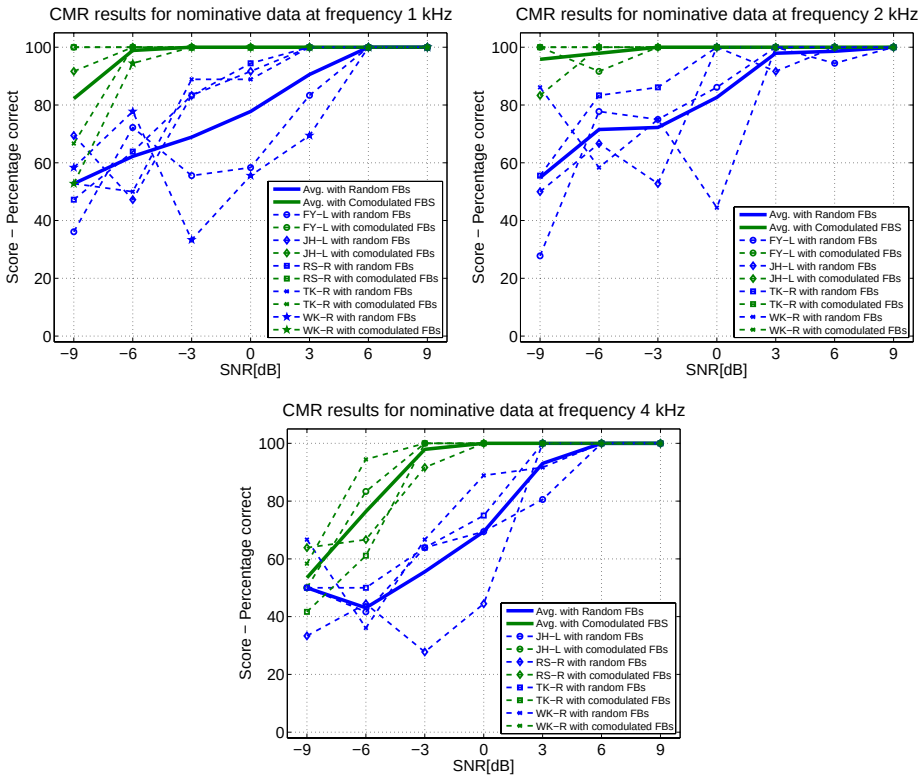


Figure 4.5: Individual and average scores: The three plots correspond to CMR results for NH listeners at 1, 2 and 4 kHz respectively. In each of the three plots, the green lines correspond to the comodulated FBs condition and the blue lines correspond to the random FBs condition. The dashed lines correspond to individual NH ear while the solid line is the average over the NH ear scores at each frequency.

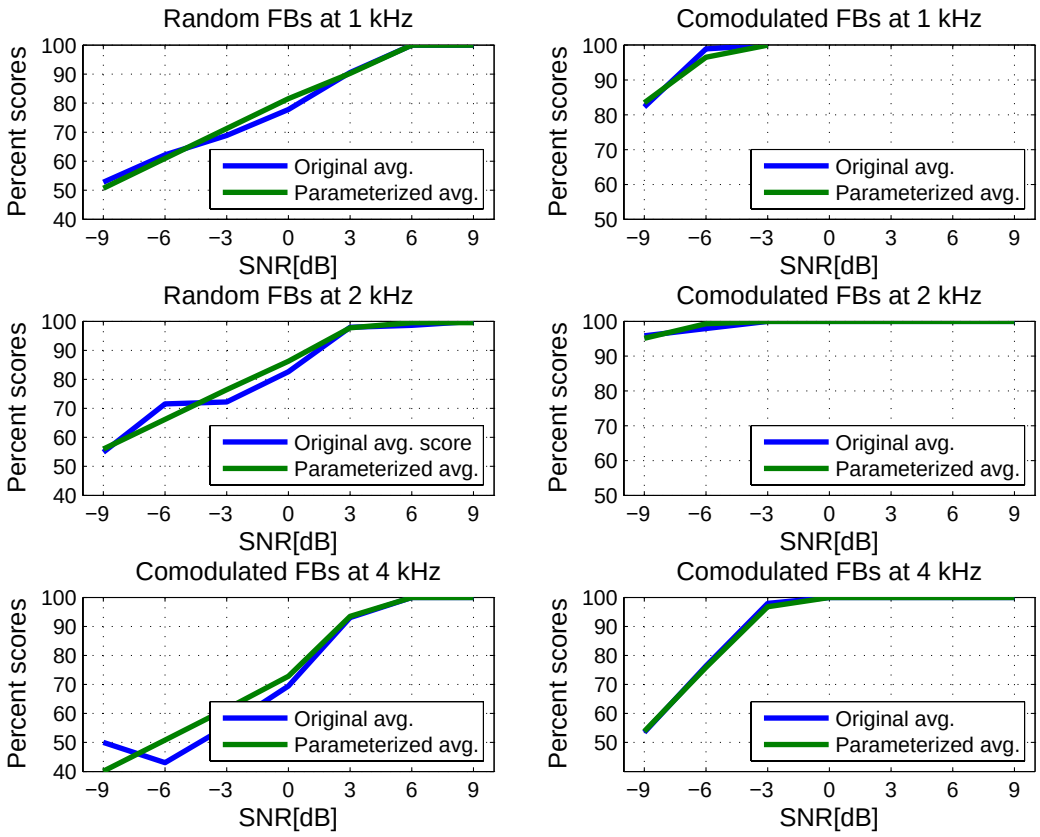


Figure 4.6: Average CMR scores across NH listeners: The three left panels correspond to the random flanking bands condition, while the three right panels correspond to the comdulated condition. The top, middle and bottom panels correspond to CMR results at 1, 2 and 4 kHz respectively. The blue line is the average NH score as a function of the SNR, while the green line is the average parameterized score. The parameterized score is obtained by fitting the actual score functions with straight lines. As seen from the figure, each of the six plots shows a good correlation between actual and parameterized data.

4.3.2 CMR results on HI ears

Figure 4.7 shows the PTC results of a normal hearing listener TK-R for 1, 2, and 4 kHz. The results indicate no dead region at these frequencies as confirmed by CMR results (Fig. 4.8), where there is a normal masking

release at these frequencies.

Figures 4.9 (for PTC) and 4.10 (for CMR) are the results with hearing-impaired ear VS-R (high frequency sloping hearing loss in the last 2 years), again tested at 1, 2, and 4 kHz. The dotted lines in the figures indicate average normal hearing data. Both PTC and CMR results indicate a dead region at 2-4 kHz. This diagnosis is confirmed by the speech perception experiment wherein listener VS-R has very low scores for CVs /fa/, /sa/, /ba/ having perceptual cues in the 2-4 kHz region [34, 40].

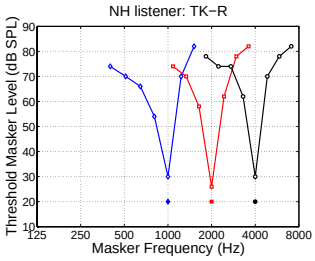


Figure 4.7: PTC results for normal hearing listener TK-R. The curves are fine tuned, which reflect good frequency selection ability of the ear. This ear has no cochlear dead regions at 1, 2 and 4 kHz.

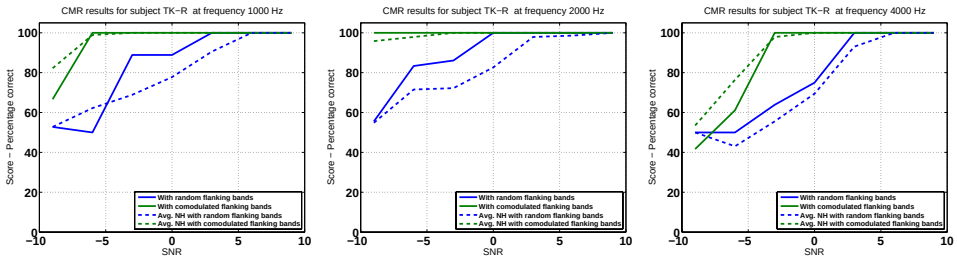


Figure 4.8: CMR results for normal hearing listener TK-R. For each of the measured frequencies of 1, 2 and 4 kHz, the ear has a higher score in the comodulated flanking bands condition as compared to the random flanking bands case. Thus, there is release of masking on comodulation, a.k.a. CMR. Again from the CMR results, the ear has no dead cochlear regions at the measured frequencies.

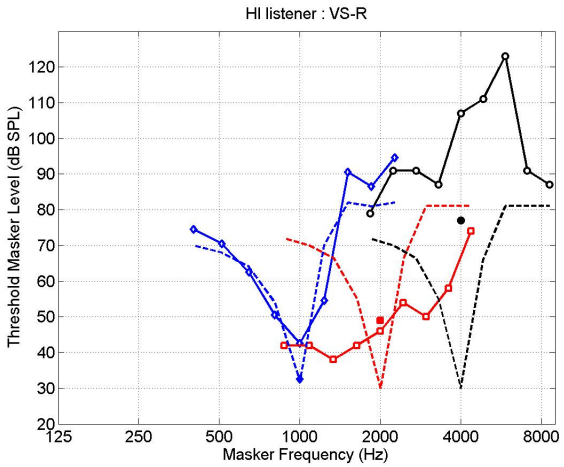


Figure 4.9: PTC results for hearing-impaired listener VS-R with normal data (in dashed lines) for reference. As seen from the figure, VS-R has good tuning at 1 kHz, which implies no CDR. However, the tuning curve is significantly shallow at 2 and 4 kHz, which is abnormal and indicative of a CDR.

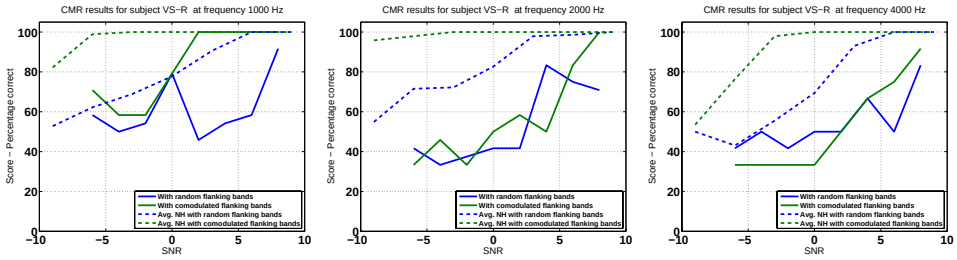


Figure 4.10: CMR results for hearing-impaired listener VS-R.

4.4 Conclusions from the CMR results

Our results appear consistent with the hypothesis that CMR can be used as a diagnostic tool to detect dead cochlear regions. The number and center frequency of the flanking bands can be conveniently chosen so that they

fall into regions where we believe there is no dead region as established by CM testing. For unilateral hearing loss, flanking bands may be introduced in the contralateral ear. Given the current experimental setup, we hope to gain insight into how a hearing-impaired ear detects speech modulations and thus forms correlations with the ability to understand speech in wideband modulations such as in gated noise [58].

CHAPTER 5

NAL-R AMPLIFICATION

Although the benefit from a hearing aid varies widely across individuals, most hearing-impaired individuals complain about receiving fewer benefits of amplification from the traditional hearing aids [37]. According to a study by [26], less than 60 percent of hearing aid users report being satisfied with their hearing aids. Despite years of such research, it remains unclear why two individuals, with the same hearing loss configuration, reveal significantly different abilities in speech understanding [66]. In particular, many listeners with mild to moderate hearing loss, who should be strong candidates for hearing aid, do not accept the effectiveness of their devices [13]. They have tried to overcome difficulties in speech understanding without using any assistive devices, which continues to limit their social ability and interaction with others.

Among various gain prescriptions for nonlinear hearing aids, variations of the *half-gain rule* have been widely used in the audiology clinic ([9], [10]). These consist of fitting the hearing aid with a gain of approximately one half the hearing threshold loss of the wearer [8]. Based upon this half-gain rule, Byrne and Harvey [9] derived the NAL-R (National Acoustic Laboratories - Revised) formula, which uses a three-frequency average gain. NAL-R is been widely used by audiologists and hearing aid specialists and is recommended for patients with mild to moderate hearing loss [14].

However, there are several problems associated with fitting hearing aids if

the NAL-R formula is directly employed using results from classic speech tests. Applying the gain prescription with either pure-tone audiometric thresholds (i.e., audiogram or pure-tone average in 0.5, 1, 2 kHz) or scores from speech recognition tests has severe limitations ([52]). This chapter presents our results of 25 HI ears on the CV discrimination task using NAL-R amplification. The methods are the same as described previously in Chapter 3 except that each presented sound was NAL-R amplified.

5.1 NAL-R fitting formula

Below is the NAL-R fitting formula used in our experiment for calculating required real-ear gain ($G_{RE}(f)$) as a function of frequency for mild to moderate hearing losses.

Step 1:

Calculate $X(dB) = 0.05 \times (HTL_{0.5} + HTL_1 + HTL_2)$ where HTL_f is the hearing threshold level of the ear at frequency f in kHz.

Step 2:

Calculate the prescribed REG at each frequency: $G_{RE}(f) = X + 0.31 \times HTL(f) + \Delta G_{RE}(f)$ where $G_{RE}(f)$ is the real-ear gain at frequency f (in kHz) while $\Delta G_{RE}(f)$ is a frequency dependent additional gain to optimize the overall loudness of the average speech spectrum. Table 5.1 shows the value of this additional gain in dB as a function of frequency.

Table 5.1: Tabulated values of the additional gain $\Delta G_{RE}(f)$ as a function of frequency (in kHz) as prescribed by the NAL-R formula.

Freq.(kHz) 0.25	0.5	1	1.5	2	3	4	6
$\Delta G_{RE}(f)$ -17	-8	-3	1	1	-1	-2	-2

5.2 Results

Next we present the results of HI performance on the CV discrimination test with and without NAL-R amplification. Figures 5.1 and 5.2 show the score vs. SNR plots for the 16 CVs for HI subject 05. The red line is the score with NAL-R amplification while the blue line represents the score without NAL-R (just flat gain). The metric that we use to quantify the difference between these two scores is the area between the two curves, which is marked in the top-let corner of each of the 16 panels. If the area is positive, it implies that NAL-R amplification improved the overall score of that particular CV; negative area implies NAL-R ameliorated scores; and an area of zero implies there was no change of score across SNR.

Subject HI_05 : BD-L

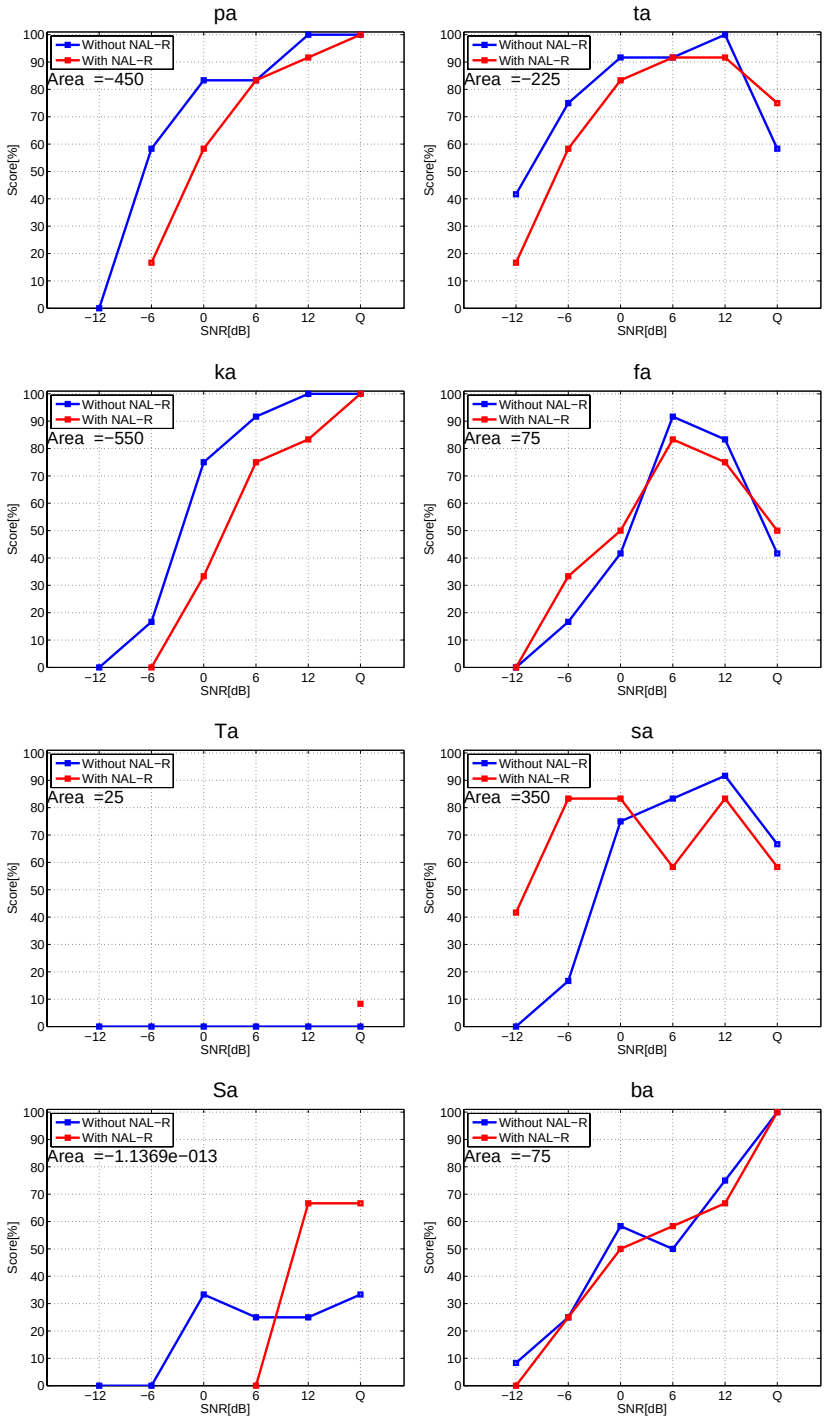


Figure 5.1: Score vs. SNR for the 8 CVs for HI subject 05.

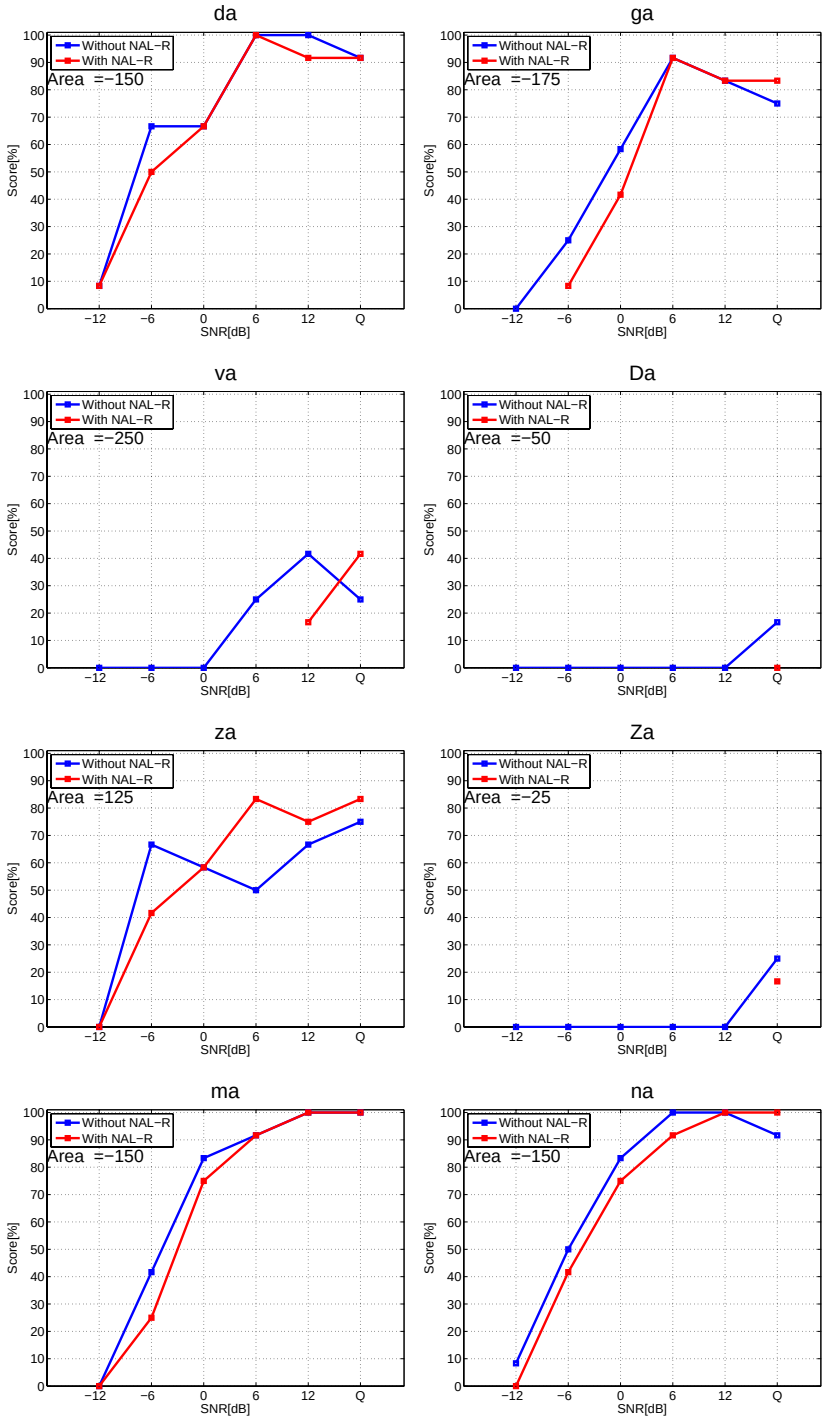


Figure 5.2: Score vs. SNR for the 8 CVs for HI subject 05.

Following is the summary of the sounds for subject 05:

- 4 sounds (fa, θa, sa, za) improved with NAL-R amplification.
- 11 sounds (pa, ta, ka, ba, da, ga, va, ða, ʒa, ma, na) deteriorated with NAL-R amplification.
- 1 sound (ʃa) remain unchanged with NAL-R amplification.

From Figs. 5.1 and 5.2, we see that only 4 sounds improved with NAL-R. We see similar results for the other 24 HI ears as well. NAL-R does not universally better scores for all CVs.

Figure 5.3 shows the bar plot for the 25 HI ears' performance on the 16 CVs with NAL-R amplification. As indicated in the legend, blue denotes the percentage of listeners who had better scores with NAL-R amplification. The percentage of listeners with degraded performance with amplification are shown in green while the brown area is the percentage of HI ears that exhibited no net change in performance on NAL-R amplification. For example, for /pa/ sound, 36% of ears (9 out of 25) had better recognition score on amplification while 60% (15 out of 25) performed poorly with amplification. One HI ear had no net change.

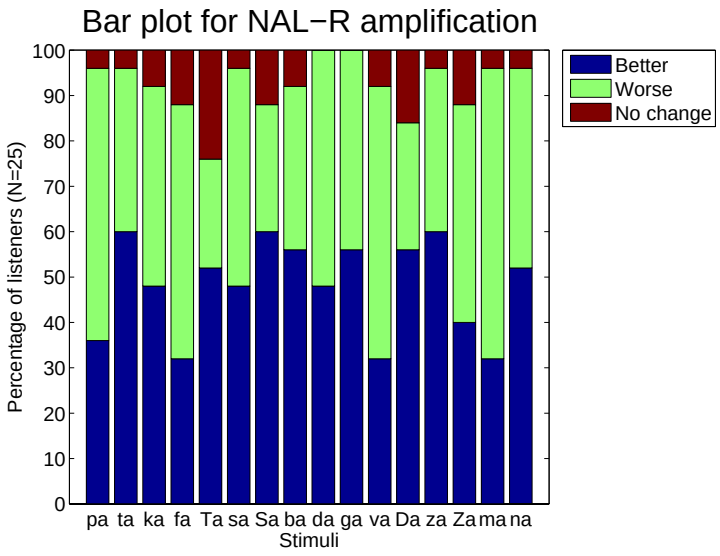


Figure 5.3: A stacked bar plot for the percentage of listeners whose scores improved/deteriorated/remain unchanged with NAL-R amplification on the 16 nonsense syllables.

5.3 Conclusions

Our results show that though NAL-R amplification may be beneficial on average, it deteriorates the perception of a few CVs. The primary reason for its failure is that it is not based on the perceptual features that are being missed by the ear. Instead, it amplifies the entire speech spectrum (depending on one's HL), which may result in amplification of conflicting cues in sound, which degrades HI perception. The presence of dead regions may also render this amplification technique useless. Based on our results, we suggest using a speech test to identify the inaudible consonants/perceptual cues for individual HI listeners, and then use a feature-based amplification scheme that compensates for the inaudible consonant cues. While we are many years away from designing such hearing aids that would be useful in

high ambient noise, we believe our research is a critical first step towards achieving this goal.

CHAPTER 6

CONCLUSIONS

From this study, the important conclusions are:

1. NH speech perception is deterministic. It is a binary decision-making task wherein you either hear the cue or not. The scores depend only on the audibility of the critical feature as shown for /t/ by [57].

2. For well articulated stop-consonants, the error is *essentially zero* for NH listeners in quiet environments as shown in Chapter 2.

3. The exponential nature of the AI model results from the distribution of scores of many tokens having different feature thresholds over an approximately 20 dB range, as suggested first by [20].

4. HI speech perception is much more complicated, since each ear may make voicing (timing) errors, or may have cochlear dead regions.

5. HI speech perception is highly consonant dependent. Only a few sounds are inaudible to each HI ear, making that ear *different* from others.

6. No speech test other than the CV discrimination test is robust to this consonant dependence.

7. SRT and PTA are poorly correlated with speech loss as measured by the nonsense syllable (maxEnt) test.

8. HI persons having a symmetrical hearing typically have asymmetric consonant confusions.

9. Comodulation masking release (CMR) provides a novel measure of cochlear dead regions (CDRs).

10. NAL-R amplification does not work uniformly across all sounds. It improves some yet degrading others. A more sensitive and personalized fitting strategy, based on selective feature amplification, is much more likely to be beneficial for hearing aids which *do no evil*.

REFERENCES

[1] Allen, J. (**1994**), "How do humans process and recognize speech?" IEEE Transactions on Speech and Audio **2**(4), 567–577.

[2] Allen, J. (**1996**), "Harvey Fletcher's role in the creation of communication acoustics," J. Acoust. Soc. Am. **99**(4), 1825–1839.

[3] Allen, J. (**2004**), "The Articulation Index is a Shannon channel capacity," in *Auditory Signal Processing: Physiology, Psychoacoustics, and Models*, edited by D. Pressnitzer, A. de Cheveign, S. McAdams, and L. Collet (Springer Verlag, New York), pp. 314–320.

[4] Allen, J. (**2005**), "Consonant recognition and the articulation index," J. Acoust. Soc. Am. **117**(4), 2212–2223.

[5] Benkí, J. (**2001**), "Place of articulation and first formant transition pattern both affect perception of voicing in English," J. Phonetics **29**, 1–22.

[6] Blumstein, S. and Stevens, K. (**1980**), "Perceptual invariance and onset spectra for stop consonants in different vowel environments," J. Acoust. Soc. Am. **67**, 648–662.

[7] Brandy, W. T. (**2002**), "Speech Audiometry," in *Handbook of Clinical Audiology*, edited by J. Katz (Lippincott Williams and Wilkins), 5th ed., pp. 96–110.

[8] Byrne, D. and Cotton, S. (**1988**), "Evaluation of the National Acoustic Laboratories' new hearing aid selection procedure," J. Speech Hear. Res. **31**, 178–186.

[9] Byrne, D. and Harvey, D. (**1986**), "The National Acoustic Laboratories' (NAL) new procedure for selecting the gain and frequency response of a hearing aid," Ear and Hearing **7**, 257–265.

[10] Byrne, D., Parkinson, A., and Newall, P. (**1990**), "Hearing aid gain and frequency response requirements for the severly/profoundly hearing impaired," Ear and Hearing **11**, 40–49.

[11] Carhart, R. (**1946**), "Speech reception in relation to pattern of pure tone loss," J. Speech Disorders **11**, 97–108.

[12] Chen, M. and Alwan, A. (**2001**), "On the perception of voicing for plosives in noise," in *Proc. EUROSPEECH,* (Proc. EUROSPEECH, Aalborg, Denmark), vol. 1.

[13] Ching, T. Y. C., Dillon, H., and Byrne, D. (**1998**), "Speech recognition of hearing-impaired listeners: Predictions from audibility and the limited role of high-frequency amplification," J. Acoust. Soc. Am. **103**, 1128–1140.

[14] Dillon, H. (**2001**), "Prescribing hearing aid performance," in *Hearing Aids* (Boomerang Press, New York), chap. 9.

[15] Dubno, J., Dirks, D., and Schaefer, A. (**1989**), "Stop-consonant recognition for normal-hearing listeners and listeners with high-frequency hearing loss. II: Articulation index predictions," J. Acoust. Soc. Am. **1**, 355–364.

[16] Festen, J. and Plomp, R. (**1986**), "Speech-reception threshold in noise with one and two hearing aids," J. Acoust. Soc. Am. **79**(2), 465–471.

[17] Fletcher, H. (**1950**), "A method of calculating hearing loss for speech from an audiogram," J. Acoust. Soc. Am. **22**, 1–5.

[18] Fletcher, H. (**1995**), *Speech and Hearing in Communication* (Acoustical Society of America, New York).

[19] Fousek, P., Svojanovsky, P., Grezl, F., and Hermansky, H. (**2004**), "New nonsense syllables database - Analyses and preliminary ASR experiments," in *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, pp. 2749–2752.

[20] French, N. and Steinberg, J. (**1947**), "Factors governing the intelligibility of speech sounds," J. Acoust. Soc. Am. **19**, 90–119.

[21] Hagerman, B. (**1984**), "Clinical measurements of speech reception threshold in noise," Scand. Audiol. **13**(1), 57–63.

[22] Hall, J., Haggard, M., and Fernandes, M. (**1984**), "Detection in noise by spectro-temporal pattern analysis," J. Acoust. Soc. Am. **76**, 50–56.

[23] Huang, J. and Hasegawa-Johnson, M. (**2008**), "Maximum mutual information estimation with unlabeled data for phonetic classification," in *Interspeech*.

[24] Humes, L., Dirks, D., Bell, T., and Ahlstbom, C. (**1986**), "Application of the articulation index and the speech transmission index to the recognition of speech by normal-hearing and hearing-impaired listeners," J. Speech Hear. Res. **29**, 447–462.

[25] Jiang, J., Chen, M., and Alwan, A. (**February 2006**), "On the perception of voicing in syllable-initial plosives in noise," J. Acoust. Soc. Am. **119**(2), 1092–1105.

[26] Kachkin, S. (**2001**), "Marketrak VI: the VA and direct mail sales spark growth in hearing aid market," Hearing Review **8**, 16–24.

[27] Killion, M. and Christensen, L. (**1998**), "The case of the missing dots: Al and SNR loss," Hearing J. **51**, 32–47.

[28] Killion, M. C. (**1997**), "SNR loss: I can hear what people say, but I can't understand them," The Hearing Review **4**(12), 8–14.

[29] Killion, M. C., Niquette, P. A., Gudmundsen, G. I., Revit, L. J., and Banerjee, S. (**2004**), "Development of a quick speech-in-noise test for measuring signal-to-noise ratio loss in normal-hearing and hearing-impaired listeners," J. Acoust. Soc. Am. **116**(4), 2395–2405.

[30] Klatt, D. and Klatt, L. (**1990**), "Analysis, synthesis, and perception of voice quality variations among male and female talkers," J. Acoust. Soc. Am. **87**, 820–857.

[31] Kujawa, S. and Liberman, C. (**November 2009**), "Adding insult to injury: Cochlear nerve degeneration after "temporary" noise-induced hearing loss," J. Neuroscience **29**(45), 14077–14085.

[32] Li, F. and Allen, J. (**2009**), "Multiband product rule and consonant identification," J. Acoust. Soc. Am. **126**(1), 347–353.

[33] Li, F. and Allen, J. (**2010**), "Manipulation of consonants in natural speech," IEEE Trans. Audio, Speech and Language Processing In press.

[34] Li, F., Menon, A., and Allen, J. (**2010**), "A psychoacoustic method to find the perceptual cues of stop consonants in natural speech," J. Acoust. Soc. Am. **127**(4), 2599–2610.

[35] Likser, L. (**1957**), "Is it VOT or a first-formant transition detector?" J. Acoust. Soc. Am. **6**, 1547–1551.

[36] Lippman, R. (**1997**), "Speech recognition by machines and humans," Speech Commun. **22**, 1–15.

[37] Margilen, G. (**1990**), *The Real Hearing Aid Market* (Hearing Center Network Publication).

[38] Markessis, E., Nasr-Addine, H., Colin, C., Hoonhorst, I., Collet, G., Deltenre, P., Munro, K. J., and Moore, B. C. (**2009**), "Effect of presentation level on diagnosis of dead regions using the threshold equalizing noise test," Int. J. Audiol. **48**(2), 55–62.

[39] Massaro, D. and Oden, G. (**March 1980**), "Evaluation and intergration of acoustic features in speech perception," J. Acoust. Soc. Am. **67**(3), 996–1013.

[40] Menon, A., Li, F., and Allen, J. (**2010**), "A new Methodology to study perceptual cues of 8 Fricative Consonants in Natural Speech," (unpublished) In preparation.

[41] Miller, G. and Nicely, P. (**1955**), "An analysis of perceptual confusions among some English consonants," J. Acoust. Soc. Am. **27**, 338–352.

[42] Moeller, M. P. (**2000**), "Early intervention and language development in children who are deaf and hard of hearing," Pediatrics **106**(3), e43.

[43] Moore, B. (**2001**), "Dead regions in the cochlea: Diagnosis, perceptual consequences, and implications for the fitting of hearing aids," Trends in Amplification **5**, 1–34.

[44] Moore, B. C. (**2004**), "Dead regions in the cochlea: Conceptual foundations, diagnosis, and clinical applications," Ear Hear **25**(2), 98–116.

[45] Moore, B. C. and Alcantara, J. I. (**2001**), "The use of psychophysical tuning curves to explore dead regions in the cochlea," Ear Hear. **22**(4), 268–278.

[46] Moore, B. C., Killen, T., and Munro, K. J. (**2003**), "Application of the TEN test to hearing-impaired teenagers with severe-to-profound hearing loss," Int. J. Audiol. **42**, 465–474.

[47] Mueller, H. G. and Killion, M. C. (**1990**), "An easy methold for calculating the articulation index," Hearing Journal **43**(9), 14–17.

[48] Patterson, R. D. and Nimmo-Smith, I. (**1980**), "Off-frequency listening and auditory-filter asymmetry," J. Acoust. Soc. Am. **67**(1), 229–245.

[49] Pavlovic, C., Studebaker, G., and Sherbecoe, R. (**1986**), "An articulation index based procedure for predicting the speech recognition performance of hearing-impaired individuals," J. Acoust. Soc. Am. **80**, 50–57.

[50] Phatak, S. and Allen, J. (**2007**), "Consonant and vowel confusions in speech-weighted noise," J. Acoust. Soc. Am. **121**(4), 2312–2326.

[51] Phatak, S., Yoon, Y., Gooler, D., and Allen, J. (**2009**), "Consonant loss profiles for hearing impaired listeners," J. Acoust. Soc. Am. **126**(5), 2683–2694.

[52] Picard, M., Banville, R., Barbarosie, T., and Manolache, M. (**1999**), "Speech audiometry in noise-exposed workers: the SRT-PTA relationship revisited," Int. J. Audiology **38**, 30–43.

[53] Plack, C. J. (**May 2005**), *Sense of Hearing* (Psychology Press), 1st ed.

[54] Plomp, R. (**1978**), "Auditory handicap of hearing impairment and the limited benefit of hearing aids," J. Acoust. Soc. Am. **63**, 533–549.

[55] Plomp, R. (**1986**), "A signal-to-noise ratio model for the speech-reception threshold of the hearing impaired," J. Speech Hear. Res. **29**, 146–154.

[56] Rankovic, C. (**1991**), "An application of the articulation index to hearing aid fitting," J. Speech Hear. Res. **34**, 391–402.

[57] Régnier, M. and Allen, J. (**2008**), "A method to identify noise-robust perceptual features: application for consonant /t/," J. Acoust. Soc. Am. **123**(5), 2801–2814.

[58] Rhebergen, K. and Versfeld, N. (**April 2005**), "A Speech Intelligibility Index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners," J. Acoust. Soc. Am. **117**(4), 2181–2192.

[59] Scharenborg, O. (**2007**), "Reaching over the gap: A review of efforts to link human and automatic speech recognition research," Speech Commun. **49**, 336–347.

[60] Shannon, C. E. (**1948**), "A mathematical theory of communication," Bell System Technical Journal **38**, 611–656.

[61] Smoorenburg, G. (**1992**), "Speech reception in quiet and in noisy conditions by individuals with noise-induced hearing loss in relation to their tone audiogram," J. Acoust. Soc. Am. **91**, 421–437.

[62] Sroka, J. and Braida, L. D. (**2005**), "Human and machine consonant recognition," Speech Commun. **45**, 401–423.

[63] Stevens, K. and Blumstein, S. (**1978**), "Invariant cues for place of articulation in stop consonants," J. Acoust. Soc. Am. **64**, 1358–1368.

[64] Sumerfield, Q. and Haggard, M. (**August 1977**), "On the dissociation of spectral and temporal cues to the voicing distinction in initial stop consonants," J. Acoust. Soc. Am. **62**(2), 435–448.

[65] Summers, V., Molis, M. R., Msch, H., Walden, B. E., Surr, R. K., and Cord, M. (**2003**), "Identifying dead regions in the cochlea: psychophysical tuning curves and tone detection in threshold-equalizing noise," Ear Hear. **24**, 133–142.

[66] Tremblay, K. L., Billings, C. J., Friesen, L. M., and Souza, P. E. (**2006**), "Neural representation of amplifed speech sounds," Ear Hear **27**, 93–103.

[67] Zurek, P. and Delhorne, L. (**1987**), "Consonant reception in noise by listeners with mild and moderate sensorineural hearing impairment," J. Acoust. Soc. Am. **82**(5), 1548–1559.

AUTHOR'S BIOGRAPHY

Riya Omprakash Singh received her B.Tech degree from the National Institute of Technology - Karnataka (NITK). She worked as a research assistant at the Beckman Institute, University of Illinois at Urbana-Champaign, from 2008 to 2010. She has held visiting student research positions at Indian Institute of Technology, Madras (IITM), and the Tata Institute of Fundamental Research, Mumbai, India, in the summer of 2006 and 2007 respectively. She also worked part-time at Mimosa Acoustics in 2009.

After she graduates in August 2010, Riya plans to move to Boston where she will join the Engineering Development Group (EDG) at The MathWorks, Inc. She can be contacted at 217-419-3696 or at riyasingh87@gmail.com.